CIS: A Deterministic, Offline-by-Default Compounding Intelligence Sandbox with Verifier-Gated Self-Improvement

Ali Uyar

2026-01-12

Workspace path convention (public-facing)

Let **WS_ROOT** = `<workspace_root>` denote the root directory of the workspace used to produce the reported results. All workspace paths in this paper are normalized as `<WS_ROOT>/...` (no machine-specific absolute paths).

Abstract

Claims of "model self-improvement" are frequently undermined by non-deterministic evaluation, informal ground truth, porous safety boundaries (e.g., network leakage, tool escapes), or mutable logs that prevent forensic auditing. We present **CIS (Compounding Intelligence Sandbox)**, an end-to-end system for **verifier-grounded** capability growth on **verifiable tasks**, designed to be **offline-by-default**, **determinism-first**, and **auditable via immutable run records**.

CIS integrates:

- A canonical CLI surface and a workspace artifact system with schema-validated, immutable run records.
- Suite snapshots with a **hidden subset** protected by **sealed evaluator mode** and explicit unlock; default evaluation excludes hidden tasks.
- A sandbox layer for tool execution with default offline enforcement and incident recording.
- Frontier governance and **difficulty indexing (DI)** calibration with **DI50** tracking.
- A self-improvement training loop combining verified distillation and **reward learning from verifiers (RLVR)** with PEFT adapters (LoRA/LoRA+/DoRA) and **promotion/rollback gating**.

We report early local results in a real workspace using **qwen2_5_coder_7b_instruct** with a **DoRA** adapter (champion: **adapter-000075**): verifier **PASS@1** on visible tasks improves from **iter-000021** to **iter-000149** across multiple suites (e.g., `core_logic` from **0.5000** to **0.8077**), while rollbacks reject regressions (e.g., `iter-000150`, `iter-000151`). We also document and validate a deterministic, training-only exploration rule for a SAT/UNSAT family (**FAMILY-004**) that addresses label collapse (policy outputs only `sat`) by restoring verified learning signal without altering evaluation behavior.

Keywords: verifiable tasks; deterministic verifiers; offline sandbox; RLVR; PEFT; provenance.

1. Introduction

Verifiable tasks—tasks paired with deterministic verifiers that accept or reject candidate outputs—provide a pragmatic substrate for measuring capability and enabling automated improvement loops. In principle, one can iteratively train a model to solve more verifiable tasks, using verifiers as the source of truth. In practice, "self-improvement" claims often fail under scrutiny for predictable reasons:

- **Non-deterministic evaluation** obscures whether measured gains are real or noise.
- **Porous safety boundaries** (e.g., network access, tool escapes) can invalidate conclusions.
- **Mutable or incomplete provenance** (missing seeds, configs, or logs) prevents reproduction.
- **Silent regressions** can be mistaken for progress unless rollback is first-class.

This paper describes **CIS**, a system intended to make verifiable-task self-improvement **reviewer-auditable**: evaluation is designed to be deterministic under explicit modes; the default execution environment is offline; artifacts and decisions are recorded in immutable run records; and promotions are gated to prevent regressions by construction.

1.1 Contributions

This report contributes the following (constrained to what is specified in the current system description and artifacts):

1. An end-to-end architecture for verifier-grounded improvement with a canonical CLI, immutable run records, and artifact-centric provenance suitable for forensic auditing.
2. Suite snapshots with hidden subsets, where hidden evaluation requires sealed evaluator mode plus an explicit unlock token; default evaluation excludes hidden tasks.
3. Offline-by-default sandboxing for tool execution with incident recording for safety violations and related anomalies.
4. A verifier-grounded training loop combining verified distillation and RLVR with PEFT adapters (LoRA/LoRA+/DoRA), coupled to promotion/rollback gating.
5. Early empirical results from a real local workspace showing PASS@1 improvements between a baseline and a champion iteration, with documented rollbacks preventing regressions.
6. A deterministic, training-only exploration rule that fixes a concrete failure mode (SAT/UNSAT label collapse in FAMILY-004) while leaving evaluation unchanged.

2. Background and Problem Setting

2.1 Verifiable tasks and verifiers

A **verifiable task** is a task for which there exists a **deterministic verifier** that maps a candidate model output to **pass/fail** (optionally with structured evidence). CIS assumes that verifiers are the source of truth: training signals (distillation targets and rewards) are grounded in verifier outcomes, not model self-assessments.

Verifier implementations are selected by `verifier_id` with deterministic backends for code (`VER-001`/`VER-002`), math (`VER-003`), logic (`VER-004`), CUT (`VER-005`..`VER-008`), and composite (`VER-011`). Each verifier emits a determinism hash, and every evaluation run performs a flakiness probe that reruns the verifier `CONST-013` times; any mismatch raises `ERR-004` and aborts the run (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).

2.2 Core terms (as used in this paper)

- **Artifact:** A schema-validated output of a CIS command, stored in the workspace artifact system. Artifacts are content-addressed by `schema_hash` computed as SHA-256 over canonical JSON (sorted keys; `schema_hash`/`created_at` removed, and content-addressed id fields omitted). `compute_schema_hash` sets both `schema_hash` and the primary id for content-addressed schemas; ART-* paths are resolved via a canonical registry (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).
- **Run record:** An immutable, schema-validated JSON record written per run, capturing provenance (seed, config hash, git state, artifact references, and promotion decisions).
- **Suite snapshot:** A fixed manifest of tasks used for evaluation and/or training, containing `task_ids` and a designated **hidden subset** (`hidden_task_ids`).
- **Hidden subset:** Tasks excluded from default evaluation; inclusion requires **sealed evaluator mode** and an explicit unlock token.
- **Sealed evaluator mode:** A mode that permits evaluation on hidden tasks only under explicit unlock; default evaluation excludes hidden tasks.
- **Deterministic mode:** An explicit CLI mode (enabled via `-deterministic`) intended to make runs repeatable under a fixed seed and controlled execution conditions.
- **Promotion / rollback:** A governance mechanism in which candidate adapters are promoted only if they pass gates (including regression checks); otherwise the system rolls back the candidate.

2.3 Metrics: PASS@1

PASS@1 denotes the fraction of evaluated tasks for which the model's **first** attempt passes the deterministic verifier.

- In CIS terminology, evaluation corresponds to `attempt_index = 0` (generation is "unchanged" relative to any training-only exploration mechanisms).
- Unless explicitly stated otherwise, PASS@1 in this report is computed on the **visible subset** of each suite snapshot (hidden tasks excluded).

Evaluation uses a single attempt per task (`attempt_index=0`) with `attempt_seed = sha256(seed|task_id)` and no generation retries; PASS@1 is the unweighted fraction of tasks whose verifier returns pass. For model decoding, `do_sample` is disabled at `attempt_index=0` (greedy decode). `max_new_tokens` comes from the task budget (default `BUDGET-001`) and is capped to 64 tokens for families `FAMILY-003..FAMILY-008`; sampling temperatures/top_p/top_k are only used when `attempt_index>0` (training-only exploration). Each evaluation also runs a verifier flakiness probe (`CONST-013` reruns); instability raises `ERR-004` and aborts the run (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).

2.4 Difficulty indexing: DI, DI50, and CI95

CIS reports difficulty-indexing summaries produced by a **DI calibration** step.

- **DI (Difficulty Index):** a scalar difficulty value produced by CIS's calibration procedure for tasks/suites.
- **DI50:** a single scalar summary reported per iteration; the name indicates a "50% point" of the DI-calibrated evaluation summary.
- **CI95:** a 95% confidence interval reported alongside DI50.

CIS fits a 1PL logistic difficulty model (Rasch-style) with MAP estimation under a zero-mean Gaussian prior (sigma=3). Task difficulty parameters `b` are derived from pass rates, and solver ability `theta` is estimated via Newton steps with an L2 penalty; DI50 is the default solver theta from the latest stable calibration. CI95 is computed as `theta +/- 1.96 * SE` from the Hessian, and stability requires median CI width <= CONST-005, bootstrap Spearman >= CONST-006, and flakiness_rate <= CONST-007 (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).

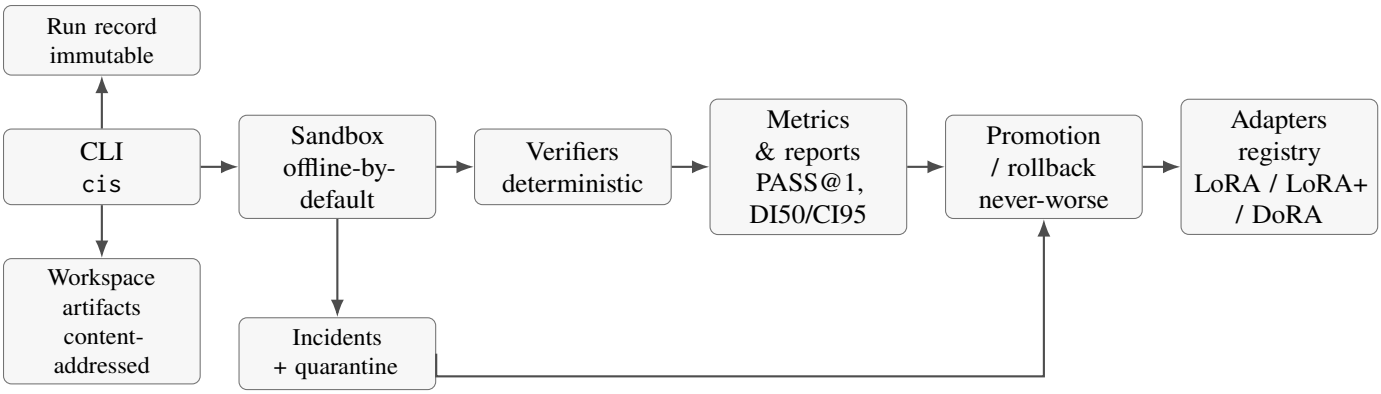

Figure 1: High-level CIS architecture. The CLI orchestrates sandboxed tool execution and deterministic verification; metrics feed promotion/rollback decisions, while provenance and incidents are recorded as immutable artifacts.

3. System Design (CIS)

3.1 High-level Architecture

At a high level, CIS organizes work around a canonical CLI (`cis`) that reads/writes workspace artifacts, routes tool execution through a sandbox, evaluates model outputs via deterministic verifiers, and governs iterative training via promotion/rollback gates. Reports are rendered from immutable artifacts and run records.

Canonical commands (as described): `init`, `run`, `eval`, `train`, `frontier govern`, `frontier calibrate`, `report render`, and `report serve`.

Full CLI help output for the commands used in this paper is provided in Appendix A.1 (captured from `python -m cis ... -help`).

3.2 Artifact & Provenance Model

CIS's provenance model is workspace-centric:

- Runs write **schema-validated artifacts** (e.g., metrics, reports) under `<WS_ROOT>/...`.
- Each run emits an immutable **run record** capturing provenance, including:
  - the run seed,
  - a configuration hash ("config hash"),
  - git state (commit and whether the repository was "dirty"), and
  - references to produced/consumed artifacts.

The system description additionally notes that run records capture **model selector** and **promotion decision** information.

Run record schema is SCH-004. Required fields include `run_id`, `created_at`, `seed`, `workspace_id`, `git`, `model_selector`, `reasoning_mode`, `suite_snapshot_ids`, `budgets`, `metrics`, `token_counts`, and `artifacts` (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md` and the redacted run records under `/evidence/runs`).

3.3 Suite Snapshots and Hidden Subsets

CIS uses **suite snapshots**: fixed manifests listing evaluated task IDs (`task_ids`) and a designated hidden subset (`hidden_task_ids`). The hidden subset is excluded from default evaluation.

- Hidden evaluation requires **sealed evaluator mode** plus an **unlock token**.
- Default evaluations and the PASS@1 results in this report are computed on the **visible** subset.

Unlock tokens are resolved to ART-041 by default and must be a non-empty file. Hidden tasks can only be evaluated when `include_hidden` + `sealed_evaluator_mode` + a valid token are all true; otherwise

evaluation raises ERR-010. Hidden-subset access is logged as a hidden_subset_access event in the workspace DB (implementation withheld; see docs/METHODS_EXTRACTED_PUBLIC.md).

3.4 Sandbox & Offline Enforcement

CIS routes tool execution through a sandbox module and enforces an offline-by-default policy:
- Network is **denied by default**, unless explicitly enabled.
- In the real workspace configuration used for the experiments reported here:
  - <WS_ROOT>/cis_config.yaml: models.allow_remote_downloads=false, sandbox.offline_default=true.
- Offline violations must raise the canonical network error and create an incident record.

CIS also includes an incidents/quarantine mechanism for safety anomalies (offline violations, verifier flakiness, determinism issues), which can trigger quarantine/rollback behavior.

Sandboxed tool execution uses Docker when available: a read-only root filesystem, /tmp tmpfs, workspace mounted at /workspace, and -network none when offline. If Docker is unavailable, a local sandbox wrapper blocks network sockets and process spawning and prevents writes outside the workspace. Tool allowlists are recorded in SCH-027 with allow_network_default=false. Incident reports are schema-validated as SCH-015 (implementation withheld; see docs/METHODS_EXTRACTED_PUBLIC.md).

3.5 Governance, DI Calibration, and Frontier Control

CIS includes a governance layer for candidate tasks and difficulty calibration:
- Candidate tasks are generated, deduplicated, scanned, and governed (the "frontier").
- DI calibration fits difficulty and tracks **DI50** with stability checks.

Frontier tasks are generated deterministically from seeded templates across multiple families (affine cut, implication, math_symbolic, logic_smt, composite, code_algo, code_bugfix) and deduplicated by normalized task hashes. Governance excludes tasks seen in the last CONST-064 training iterations, applies saturation retirement, enforces DI variance near CONST-024, and emits calibration/sentinel/alert buckets in the ART-032 candidate set manifest (implementation withheld; see docs/METHODS_EXTRACTED_PUBLIC.md).

3.6 Training Loop: Distillation + RLVR + PEFT

CIS implements an iterative training loop grounded in verifiers:
1. **Verified distillation:** collect samples that pass deterministic verifiers and distill them into the policy.
2. **RLVR (reward learning from verifiers):** compute verifier-grounded rewards and train the policy accordingly.
3. **PEFT training:** train parameter-efficient adapters (LoRA/LoRA+/DoRA).
4. **Gating:** validate and decide whether to promote or roll back the new adapter.

This paper reports results using DoRA adapters (PEFT) on the base model **qwen2_5_coder_7b_instruct** (see Section 5).

RLVR datasets are built by sampling visible tasks from stable suites plus DI-band frontier tasks, generating groups of candidates per task, assigning binary rewards from verifier pass/fail, and standardizing advantages within each group (implementation withheld; see docs/METHODS_EXTRACTED_PUBLIC.md). **TODO:** Provide the optimizer/hyperparameters and training schedule for the reported runs.

3.7 Promotion/Rollback Gates (Never-Worse)

CIS treats regression prevention as a primary system requirement:
- Candidate adapters are evaluated against previously passed tasks and broader suite metrics.
- Adapters that regress on previously passed tasks are rejected and rolled back.
- The system records rollback decisions in iteration reports and run records (examples in Section 6).

Promotion uses a "never-worse" regression gate over stable suites (`SUITE-001..SUITE-004`): any task passed by the baseline but failed by the challenger counts as a regression; promotion requires `regression_count=0`. Decisions also require adapter-check pass (`ERR-011` on failure), teacher hint dependence pass (`ERR-017`), and ladder regression pass (`ERR-018`), and they promote only if max PASS@1 delta >= CONST-016 or DI50 delta >= CONST-017 (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).

4. Experimental Setup

All reported results are produced in a "real model" workspace under `<WS_ROOT>`.

4.1 Model and adapters

- **Base model:** `qwen2_5_coder_7b_instruct`
- **Champion adapter (time of writing):** `adapter-000075`
- **Adapters registry:** `<WS_ROOT>/registries/adapters.json`

The champion adapter's PEFT configuration is recorded alongside its checkpoint:
- `<WS_ROOT>/adapters/adapter-000075/checkpoint/adapter_config.json`
- `peft_method=dora`, `rank=16`, `alpha=32`, `dropout=0.0`

4.2 Suites, snapshots, and hidden subsets

Suite snapshots used in this report include a hidden subset excluded from default evaluation. PASS@1 values reported in Section 6 are computed on the **visible subset**.

4.3 Execution environment

CIS is implemented in **Python 3.11**.

TODO: Provide full environment fingerprints required for strict reproduction (OS, Python patch version, GPU model, driver/CUDA versions, container image hash if applicable, and all key dependency versions beyond the manifests in `pyproject.toml` and `docker/sandbox_base.Dockerfile`).

4.4 Evaluation protocol (as currently specified)

- Reported PASS@1 values are taken from iteration metrics artifacts:
  - Baseline: `<WS_ROOT>/reports/iter-000021/metrics.json`
  - Champion: `<WS_ROOT>/reports/iter-000149/metrics.json`
- PASS@1 is computed using deterministic verifiers on the visible subset of each suite snapshot.

See Section 3.3 for decoding configuration, attempt budgeting, and PASS@1 aggregation details with in-repo references.

Table 1: Suite snapshots used for the reported results. Default evaluation excludes hidden tasks; sealed evaluator mode plus unlock is required to include them.

Suite	Snapshot	Total tasks	Hidden tasks	Visible tasks
core_code	60b03d...	8	1	7
core_math	a6251d...	10	2	8
core_logic	9b2bba...	64	12	52
concept_understanding	90f4fc...	32	6	26

5. Results

5.1 Baseline vs. champion PASS@1

We compare an early baseline iteration against the current champion iteration:
- Baseline metrics: <WS_ROOT>/reports/iter-000021/metrics.json
- Champion metrics: <WS_ROOT>/reports/iter-000149/metrics.json
  The champion promotion associated with the iter-000149 results is recorded as:
- Run record: <WS_ROOT>/runs/run-20260112T124809Z-d74cd05c/run_record.json (seed 52010)
- Report: <WS_ROOT>/reports/iter-000149/summary.md

5.2 DI50 tracking (difficulty indexing)

For iter-000149, the report includes:
- **DI50:** 2.2083 with **CI95** [1.7033, 2.7133] (see <WS_ROOT>/reports/iter-000149/summary.md)
  See Section 3.4 for the DI calibration model and CI95 definition.

5.3 Rollbacks prevent regressions

CIS rejects challenger adapters that regress on previously passed tasks. The following rollbacks are recorded:
- iter-000150: rollback of adapter-000076 due to regression on logic_sat_48
  - Report: <WS_ROOT>/reports/iter-000150/summary.md
  - Run record: <WS_ROOT>/runs/run-20260112T131450Z-ec2875b8/run_record.json
- iter-000151: rollback of adapter-000077 due to regressions on logic_sat_10 and logic_sat_44
  - Report: <WS_ROOT>/reports/iter-000151/summary.md
  - Run record: <WS_ROOT>/runs/run-20260112T142709Z-9477cd3c/run_record.json

Interpretation (conservative): These records demonstrate that CIS's governance workflow can detect and reject regressions on named tasks and document the decision in artifacts. Implementation details are withheld; broader generalization beyond these suites is not claimed.

Table 2: PASS@1 on visible tasks for suite snapshots in Table 1. PASS@1 uses deterministic verifiers with `attempt_index=0` and per-task budgets; decoding configuration and aggregation are specified in Section 3.3.

Suite	Baseline PASS@1 (iter-000021)	Champion PASS@1 (iter-000149)
`core_code`	0.0000	1.0000
`core_math`	0.1250	1.0000
`core_logic`	0.5000	0.8077
`concept_understanding`	0.0000	1.0000

6. Debugged Failure Mode: SAT/UNSAT Label Collapse

6.1 Symptom

Within a SAT/UNSAT family in `core_logic`, the policy collapsed to always output `sat`. This caused two coupled failures:

- Verified distillation collected **zero verified UNSAT** samples for some tasks.
- RLVR produced **all-zero rewards** (no verified successes), yielding a training plateau.

6.2 Fix: deterministic training-only exploration for FAMILY-004

To restore learning signal without contaminating evaluation, CIS applies a deterministic exploration policy **only for training attempts** on the logic family (FAMILY-004):

- For `attempt_index > 0` (training attempts only), deterministically flip `sat` ↔ `unsat` on odd attempts.
- For `attempt_index = 0` (evaluation), generation is unchanged.

Implementation note: the mechanism is implemented and exercised in the private codebase; details are withheld in this docs-only release.

Empirically, this restored verified UNSAT samples in distillation datasets and unlocked promotions (as referenced by the system reports; see Section 6).

TODO: Quantify the effect with explicit before/after dataset statistics and/or learning curves (without adding new, unverified results).

7. Reproducibility Notes

7.1 Typical commands used

Commands invoked in the real workspace include:

- Train: `./cisq.ps1 train -deterministic -seed 52010 -json`
- Render report: `./cisq.ps1 report render -iter-id latest -json`

7.2 Dependency manifest (repo)

The repository declares runtime dependencies in `pyproject.toml` (Python >=3.11) including numpy, pydantic, PyYAML, jsonschema, torch, safetensors, sympy, z3-solver, and pyarrow; test dependencies include pytest. The sandbox base image pins pytest==8.4.0, sympy==1.13.3, and z3-solver==4.15.3.0 in `docker/sandbox_base.Dockerfile`.

Public methods summary: see `docs/METHODS_EXTRACTED_PUBLIC.md` for a redacted, paper-linked evidence map; internal implementation references are available upon request.

7.3 Determinism contract (reviewer-facing)

CIS exposes a **deterministic mode** intended to make runs repeatable. Based on the current system description and artifacts, the following contract is supported:

1. **Mode control:** Determinism is an explicit CLI mode enabled via `-deterministic` for training and evaluation runs.
2. **Seed control:** A run seed is provided via `-seed <seed>` and is recorded in the run record (e.g., seed `52010` for the promotion run in Section 6).
3. **Evaluation attempt semantics:** Evaluation behavior corresponds to `attempt_index = 0`; training-only exploration rules (Section 7) apply only when `attempt_index > 0`.
4. **Offline coupling:** In the reported workspace configuration, `sandbox.offline_default=true` and `models.allow_remote_downloads=false` in `<WS_ROOT>/cis_config.yaml`, preventing network-dependent variability by default.
5. **Sandbox mediation:** Tool execution is routed through the sandbox module; offline mode is enforced via Docker `-network none` or local socket blocking, and the spec mandates ERR-002 + incident on network attempts (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).
6. **Immutable provenance:** Each run writes an immutable, schema-validated run record capturing at least: seed, config hash, git state (including commit and "dirty" status), artifact references, model selector, and promotion decision information.
7. **Config hashing:** The config hash is SHA-256 over the resolved config mapping serialized as canonical JSON (sorted keys) (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).
8. **Artifact addressing:** `compute_schema_hash` computes SHA-256 over canonical JSON (sorted keys, `schema_hash`/`created_at` removed; primary id removed for content-addressed schemas) and writes both `schema_hash` and the primary id; ART-* paths are resolved via the canonical registry (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).
9. **Environment fingerprinting requirement:** Hardware and library nondeterminism (e.g., GPU kernel nondeterminism) is not addressed in the current draft; strict reproducibility therefore additionally requires matching environment fingerprints. **TODO:** report OS/GPU/driver/CUDA/container and dependency versions for the runs reported here.
10. **Incident handling for determinism:** `CMD-013` (`cis inspect determinism-check`) reruns evaluation for `-runs N` and compares pass vectors; mismatches set ERR-001 and mark the run as failed in the run record (implementation withheld; see `docs/METHODS_EXTRACTED_PUBLIC.md`).

7.4 Provenance pointers

An example run record (promotion run cited in Section 6) is located at:

- `<WS_ROOT>/runs/run-20260112T124809Z-d74cd05c/run_record.json`

The system description states that run records capture git commit and whether the repository was "dirty" at run time; strict reproduction should therefore prefer reruns from a clean commit with identical seeds.

8. Limitations

1. **Scope and scale:** The reported suite is small and designed for fast local iteration; these results do not establish broad generalization.
2. **Hidden subset not evaluated here:** Default PASS@1 excludes hidden subsets; sealed evaluator mode exists but is not used for the reported PASS@1 values.
3. **Incomplete methodological specification:** Remaining gaps are limited to environment fingerprints and optimizer/hyperparameter disclosure for the reported runs; these are enumerated as TODOs in Sections 4.6 and 8.
4. **No external baselines:** The reported gains are relative to an internal baseline iteration (iter-000021) and

do not compare to external benchmarks or competing methods.

5. **Security model scope:** The threat model and offline policy are implemented in the private codebase; remaining work is documenting the runtime environment for the reported runs.

9. Related Work

CIS touches several well-established areas:

- **Verifier-based learning and evaluation:** using unit tests or deterministic checkers as ground truth for training and scoring.
- **Program synthesis and tool-assisted reasoning with verifiers:** where verifiers constrain solution validity.
- **Difficulty estimation / IRT-style calibration:** using fitted difficulty indices to track evaluation frontier movement.
- **PEFT methods:** LoRA, LoRA+, and DoRA for adapter-based fine-tuning.

TODO: Add formal citations for verifier-based learning, IRT/difficulty estimation, and PEFT (LoRA/LoRA+/DoRA), as well as any RLVR-related prior work relevant to "reward learning from verifiers."

10. Conclusion

This paper presents **CIS**, a deterministic, offline-by-default, artifact-centric system for verifier-grounded iterative improvement on verifiable tasks. CIS emphasizes immutable provenance (run records), sandboxed tool execution with incident recording, suite snapshots with hidden subsets protected by sealed evaluator mode, and promotion/rollback gating designed to prevent regressions.

Early local results in a real workspace using `qwen2_5_coder_7b_instruct` with a DoRA adapter show PASS@1 improvements across multiple suites from `iter-000021` to `iter-000149`, while rollback records demonstrate regression rejection on named tasks. We also document a concrete failure mode—SAT/UNSAT label collapse in FAMILY-004—and a deterministic, training-only exploration rule that restores verified learning signal without altering evaluation behavior.

A. Schemas, CLI Summary, and Artifact Index (optional)

A.1 CLI surface (as described)

- `cis init`
- `cis run`
- `cis eval`
- `cis train`
- `cis frontier govern`
- `cis frontier calibrate`
- `cis report render`
- `cis report serve`

CLI help output captured from local `python -m cis ... -help`:

`cis -help`:

```
usage: cis [-h]
           {artifact,concept,demo,eval,frontier,init,inspect,kill,ladder,longctx,report,run,sandbox,s⌋
           ↪  cale,teacher,train,ttt}
           ...

positional arguments:
  {artifact,concept,demo,eval,frontier,init,inspect,kill,ladder,longctx,report,run,sandbox,scale,tea⌋
  ↪  cher,train,ttt}

options:
  -h, --help            show this help message and exit
```

`cis init -help`:

```
usage: cis init [-h] [--deterministic] [--allow-network]
                [--workspace WORKSPACE] [--seed SEED] [--config CONFIG]
                [--model MODEL] [--adapter ADAPTER] [--suite SUITE]
                [--budget BUDGET] [--json] [--strict] [--unsafe-experimental]
                [--unlock-token-path UNLOCK_TOKEN_PATH] [--runs RUNS]
                [--task-id TASK_ID] [--iter-id ITER_ID]
                [--parallel-jobs PARALLEL_JOBS]
                [--capacity-setting CAPACITY_SETTING]
                [--compute-budgets COMPUTE_BUDGETS]
                [--artifact-id ARTIFACT_ID] [--peft-method PEFT_METHOD]
                [--reasoning-mode REASONING_MODE]
                [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                [--quiet-debug] [--teacher-package-id TEACHER_PACKAGE_ID]
                [--allow-holdout-ttt]
                [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
```

```
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis run -help`:

```
usage: cis run [-h] [--deterministic] [--allow-network]
               [--workspace WORKSPACE] [--seed SEED] [--config CONFIG]
               [--model MODEL] [--adapter ADAPTER] [--suite SUITE]
               [--budget BUDGET] [--json] [--strict] [--unsafe-experimental]
               [--unlock-token-path UNLOCK_TOKEN_PATH] [--runs RUNS]
               [--task-id TASK_ID] [--iter-id ITER_ID]
               [--parallel-jobs PARALLEL_JOBS]
               [--capacity-setting CAPACITY_SETTING]
               [--compute-budgets COMPUTE_BUDGETS] [--artifact-id ARTIFACT_ID]
               [--peft-method PEFT_METHOD] [--reasoning-mode REASONING_MODE]
               [--baseline-reasoning-mode BASELINE_REASONING_MODE]
               [--quiet-debug] [--teacher-package-id TEACHER_PACKAGE_ID]
               [--allow-holdout-ttt]
               [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
```

```
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis eval -help`:

```
usage: cis eval [-h] [--deterministic] [--allow-network]
                [--workspace WORKSPACE] [--seed SEED] [--config CONFIG]
                [--model MODEL] [--adapter ADAPTER] [--suite SUITE]
                [--budget BUDGET] [--json] [--strict] [--unsafe-experimental]
                [--unlock-token-path UNLOCK_TOKEN_PATH] [--runs RUNS]
                [--task-id TASK_ID] [--iter-id ITER_ID]
                [--parallel-jobs PARALLEL_JOBS]
                [--capacity-setting CAPACITY_SETTING]
                [--compute-budgets COMPUTE_BUDGETS]
                [--artifact-id ARTIFACT_ID] [--peft-method PEFT_METHOD]
                [--reasoning-mode REASONING_MODE]
                [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                [--quiet-debug] [--teacher-package-id TEACHER_PACKAGE_ID]
                [--allow-holdout-ttt]
                [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis train -help`:

```
usage: cis train [-h] [--deterministic] [--allow-network]
```

```
                [--workspace WORKSPACE] [--seed SEED] [--config CONFIG]
                [--model MODEL] [--adapter ADAPTER] [--suite SUITE]
                [--budget BUDGET] [--json] [--strict] [--unsafe-experimental]
                [--unlock-token-path UNLOCK_TOKEN_PATH] [--runs RUNS]
                [--task-id TASK_ID] [--iter-id ITER_ID]
                [--parallel-jobs PARALLEL_JOBS]
                [--capacity-setting CAPACITY_SETTING]
                [--compute-budgets COMPUTE_BUDGETS]
                [--artifact-id ARTIFACT_ID] [--peft-method PEFT_METHOD]
                [--reasoning-mode REASONING_MODE]
                [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                [--quiet-debug] [--teacher-package-id TEACHER_PACKAGE_ID]
                [--allow-holdout-ttt]
                [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis frontier -help`:

```
usage: cis frontier [-h] {calibrate,govern,plateau-check} ...

positional arguments:
  {calibrate,govern,plateau-check}

options:
  -h, --help            show this help message and exit
```

`cis frontier govern -help`:

```
usage: cis frontier govern [-h] [--deterministic] [--allow-network]
```

```
                          [--workspace WORKSPACE] [--seed SEED]
                          [--config CONFIG] [--model MODEL]
                          [--adapter ADAPTER] [--suite SUITE]
                          [--budget BUDGET] [--json] [--strict]
                          [--unsafe-experimental]
                          [--unlock-token-path UNLOCK_TOKEN_PATH]
                          [--runs RUNS] [--task-id TASK_ID]
                          [--iter-id ITER_ID] [--parallel-jobs PARALLEL_JOBS]
                          [--capacity-setting CAPACITY_SETTING]
                          [--compute-budgets COMPUTE_BUDGETS]
                          [--artifact-id ARTIFACT_ID]
                          [--peft-method PEFT_METHOD]
                          [--reasoning-mode REASONING_MODE]
                          [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                          [--quiet-debug]
                          [--teacher-package-id TEACHER_PACKAGE_ID]
                          [--allow-holdout-ttt]
                          [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis frontier calibrate -help`:

```
usage: cis frontier calibrate [-h] [--deterministic] [--allow-network]
                              [--workspace WORKSPACE] [--seed SEED]
                              [--config CONFIG] [--model MODEL]
                              [--adapter ADAPTER] [--suite SUITE]
                              [--budget BUDGET] [--json] [--strict]
                              [--unsafe-experimental]
                              [--unlock-token-path UNLOCK_TOKEN_PATH]
                              [--runs RUNS] [--task-id TASK_ID]
```

```
                            [--iter-id ITER_ID]
                            [--parallel-jobs PARALLEL_JOBS]
                            [--capacity-setting CAPACITY_SETTING]
                            [--compute-budgets COMPUTE_BUDGETS]
                            [--artifact-id ARTIFACT_ID]
                            [--peft-method PEFT_METHOD]
                            [--reasoning-mode REASONING_MODE]
                            [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                            [--quiet-debug]
                            [--teacher-package-id TEACHER_PACKAGE_ID]
                            [--allow-holdout-ttt]
                            [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis report -help`:

```
usage: cis report [-h] {render,serve} ...

positional arguments:
  {render,serve}

options:
  -h, --help      show this help message and exit
```

`cis report render -help`:

```
usage: cis report render [-h] [--deterministic] [--allow-network]
                         [--workspace WORKSPACE] [--seed SEED]
```

```
                         [--config CONFIG] [--model MODEL] [--adapter ADAPTER]
                         [--suite SUITE] [--budget BUDGET] [--json] [--strict]
                         [--unsafe-experimental]
                         [--unlock-token-path UNLOCK_TOKEN_PATH] [--runs RUNS]
                         [--task-id TASK_ID] [--iter-id ITER_ID]
                         [--parallel-jobs PARALLEL_JOBS]
                         [--capacity-setting CAPACITY_SETTING]
                         [--compute-budgets COMPUTE_BUDGETS]
                         [--artifact-id ARTIFACT_ID]
                         [--peft-method PEFT_METHOD]
                         [--reasoning-mode REASONING_MODE]
                         [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                         [--quiet-debug]
                         [--teacher-package-id TEACHER_PACKAGE_ID]
                         [--allow-holdout-ttt]
                         [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

`cis report serve -help`:

```
usage: cis report serve [-h] [--deterministic] [--allow-network]
                        [--workspace WORKSPACE] [--seed SEED]
                        [--config CONFIG] [--model MODEL] [--adapter ADAPTER]
                        [--suite SUITE] [--budget BUDGET] [--json] [--strict]
                        [--unsafe-experimental]
                        [--unlock-token-path UNLOCK_TOKEN_PATH] [--runs RUNS]
                        [--task-id TASK_ID] [--iter-id ITER_ID]
                        [--parallel-jobs PARALLEL_JOBS]
                        [--capacity-setting CAPACITY_SETTING]
                        [--compute-budgets COMPUTE_BUDGETS]
```

```
                        [--artifact-id ARTIFACT_ID]
                        [--peft-method PEFT_METHOD]
                        [--reasoning-mode REASONING_MODE]
                        [--baseline-reasoning-mode BASELINE_REASONING_MODE]
                        [--quiet-debug]
                        [--teacher-package-id TEACHER_PACKAGE_ID]
                        [--allow-holdout-ttt]
                        [--holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH]

options:
  -h, --help            show this help message and exit
  --deterministic
  --allow-network
  --workspace WORKSPACE
  --seed SEED
  --config CONFIG
  --model MODEL
  --adapter ADAPTER
  --suite SUITE
  --budget BUDGET
  --json
  --strict
  --unsafe-experimental
  --unlock-token-path UNLOCK_TOKEN_PATH
  --runs RUNS
  --task-id TASK_ID
  --iter-id ITER_ID
  --parallel-jobs PARALLEL_JOBS
  --capacity-setting CAPACITY_SETTING
  --compute-budgets COMPUTE_BUDGETS
  --artifact-id ARTIFACT_ID
  --peft-method PEFT_METHOD
  --reasoning-mode REASONING_MODE
  --baseline-reasoning-mode BASELINE_REASONING_MODE
  --quiet-debug
  --teacher-package-id TEACHER_PACKAGE_ID
  --allow-holdout-ttt
  --holdout-ttt-unlock-token-path HOLDOUT_TTT_UNLOCK_TOKEN_PATH
```

A.2 Key artifact locations referenced in this paper

- Config: <WS_ROOT>/cis_config.yaml
- Adapters registry: <WS_ROOT>/registries/adapters.json
- Champion adapter config: <WS_ROOT>/adapters/adapter-000075/checkpoint/adapter_config.json
- Baseline metrics: <WS_ROOT>/reports/iter-000021/metrics.json
- Champion metrics: <WS_ROOT>/reports/iter-000149/metrics.json
- Champion report summary: <WS_ROOT>/reports/iter-000149/summary.md
- Promotion run record (seed 52010): <WS_ROOT>/runs/run-20260112T124809Z-d74cd05c/run_record.json
- Rollback reports: <WS_ROOT>/reports/iter-000150/summary.md, <WS_ROOT>/reports/iter-000151/summary.md
- Rollback run records: <WS_ROOT>/runs/run-20260112T131450Z-ec2875b8/run_record.json, <WS_ROOT>/runs/run-20260112T142709Z-9477cd3c/run_record.json

Public evidence note: redacted copies of referenced artifacts are provided under /evidence in this docs-only release (when available).