

# GOVERNED SELF-IMPROVEMENT FOR LOGICAL REASONING: EDIT-TIME GOVERNANCE FOR DEVELOPMENTAL CONSISTENCY

**Anonymous authors**

Paper under double-blind review

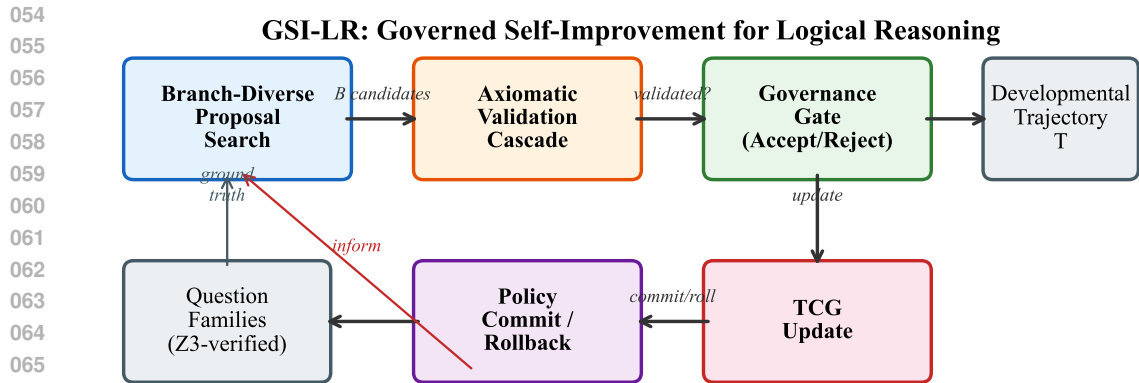
## ABSTRACT

Self-refinement methods enable large language models to improve without retraining, yet they optimize local answers rather than the future reasoner. In logical reasoning, every answer creates *longitudinal commitments*: paraphrases, negations, implication chains, and premise permutations must remain jointly consistent across developmental time. We present a governance-oriented framework and evaluation lens with proof-of-concept validation on a controlled propositional-logic domain. (1) We frame self-improvement as a *commitment-management* problem and show that uncontrolled search can *increase* contradictions even while raising accuracy. (2) We propose GSI-LR (Governed Self-Improvement for Logical Reasoning), a framework combining branch-diverse proposal search, a temporal contradiction graph (TCG) grounded in AGM-style belief revision, an axiomatic validation cascade using symbolic solvers at *edit time*, and an explicit edit-rights policy. (3) We introduce Developmental Consistency Evaluation (DCE), a protocol measuring family contradiction rate (FCR; lower is better — fewer family contradictions), acceptance precision, delayed regression, rollback burden, and maintenance debt over trajectories rather than snapshots. (4) We validate GSI-LR on a Z3-grounded propositional-logic domain (200 questions, 40 families, 50 edit rounds, 5 seeds), demonstrating that governed development occupies a favorable position on the accuracy–consistency Pareto frontier: it reduces FCR by 8.8% relative to static baselines (FCR 0.675 vs. 0.740, lower is better) while maintaining strict non-regression, whereas unconstrained search achieves perfect accuracy at the cost of *increased* contradictions (FCR 0.775).

## 1 INTRODUCTION

Large language models can decompose problems, generate multi-step explanations, and sometimes repair their own mistakes (Bubeck et al., 2023; Guo et al., 2025; Wei et al., 2026). Yet logical reasoning remains brittle. Models often answer an isolated question correctly but fail on a closely related variant, reverse a conclusion under innocuous paraphrase, or emit answers that cannot be jointly maintained (Cheng et al., 2025; Plaata et al., 2024; Xu et al., 2025; Liu et al., 2025). Recent work has shown that even advanced reasoning models can exhibit an “illusion of thinking” where surface-level chain-of-thought masks fragile underlying reasoning (Shojaee et al., 2025; Song et al., 2026). In science, law, and policy, local fluency is insufficient when global commitments are unstable.

**Problem.** Consider a concrete failure. A system correctly infers “All mammals breathe” and answers “Yes” to “Do whales breathe?” It then undergoes self-improvement: a critic identifies an error on a cetacean classification question, and a local patch modifies the system’s handling of marine animals. After the edit, the system now answers “No” to “Do whales breathe?” — contradicting its unchanged commitment about mammals. The edit improved one answer while silently breaking a *family* of related commitments. This failure is invisible to per-instance accuracy but devastating for a deployed reasoner.



077  
078  
079  
080  
081  
082  
083

Figure 1: GSI-LR framework architecture. The main loop cycles through branch-diverse proposal, axiomatic validation, governance gate, TCG update, and policy commit/rollback. Question families with Z3-verified ground truth provide the evaluation substrate. The TCG feeds diagnostic evidence back into proposal search (red arrow), closing the governance loop.

084  
085  
086  
087  
088  
089  
090

**Gap.** Existing self-refinement methods — Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), STaR (Zelikman et al., 2022) — optimize *local* answer quality. Reward modeling for RL-based reasoning (Pan et al., 2026) and recursive decomposition approaches (Qasim et al., 2025) similarly target per-instance performance. Solver-aided methods — Logic-LM (Pan et al., 2023), LINC (Olausson et al., 2023) — use solvers at *answer time*. Neither maintains longitudinal contradiction memory nor governs which self-edits survive into the future reasoner.

091  
092  
093  
094  
095  
096  
097  
098  
099

**Thesis.** We argue that the right optimization target is not the single answer but the *reasoning policy under longitudinal commitment*. We present GSI-LR as a framework paper with controlled validation on a deterministic propositional-logic domain — not a claim of state-of-the-art on standard reasoning benchmarks, but a proof-of-concept demonstrating that governed development occupies a favorable position on the accuracy–consistency Pareto frontier.

100  
101  
102

**Contributions.** Our primary contributions are conceptual and evaluative:

1. **Commitment-management framing:** we recast self-improvement as a longitudinal consistency problem and show that optimizing local accuracy can *worsen* global coherence.
2. **Developmental Consistency Evaluation (DCE):** a trajectory-level evaluation protocol with seven metrics — FCR, acceptance precision, delayed regression rate, rollback burden, maintenance debt, refusal calibration, and accuracy — that surface failure modes invisible to snapshot evaluation.

103  
104  
105  
106  
107

As an exploratory proof-of-concept, we instantiate this framing in GSI-LR, a governed edit-time pipeline combining:

3. **Branch-diverse proposal search, a temporal contradiction graph (TCG)** grounded in AGM belief revision, an **axiomatic validation cascade** using symbolic solvers at edit time, and an **edit-rights governance** policy.
4. **Controlled validation** on a Z3-grounded propositional domain (200 questions, 40 families, 50 rounds, 5 seeds) confirming the accuracy–consistency Pareto tradeoff: governed development achieves the lowest FCR (0.675 vs. 0.740 static, lower is better) while unconstrained search increases contradictions (FCR 0.775) despite perfect accuracy.

Full details appear in Appendices A–E.

## 2 RELATED WORK

**Self-refinement and critique-correct loops.** Self-Refine (Madaan et al., 2023) demonstrates iterative self-feedback without additional training. Reflexion (Shinn et al., 2023) adds verbal reinforcement signals across episodes. Self-RAG (Asai et al., 2024) trains models to retrieve and self-reflect on demand. DiffCoT (Cao et al., 2026) applies diffusion-style iterative denoising to chain-of-thought reasoning. These methods optimize *local* answer quality. They do not maintain longitudinal contradiction memory, do not govern which edits survive into the future reasoner, and do not measure family-level consistency. GSI-LR addresses this gap by adding temporal memory and governed admission.

**Solver-aided reasoning.** Logic-LM (Pan et al., 2023) connects LLMs to symbolic solvers for faithful reasoning. LINC (Olausson et al., 2023) uses first-order logic as an intermediate representation. Faithful Chain-of-Thought (Lyu et al., 2023) decomposes problems into symbolic subtasks. Thatikonda et al. (Thatikonda et al., 2026) improve symbolic translation quality for logical reasoning. Selection-Inference (Creswell et al., 2022) decomposes reasoning into alternating selection and inference steps, while Creswell and Shanahan (Creswell and Shanahan, 2022) enforce faithfulness through structured generation. Neural-symbolic integration (H6u, 2025; Garc3ez and Lamb, 2023) aims to combine the flexibility of neural models with the guarantees of symbolic systems. These approaches use solvers at *answer time* to improve current predictions. GSI-LR uses solvers at *edit time* — a fundamentally different role where the solver governs which self-modifications become permanent.

**Recursive self-improvement.** The Darwin Gödel Machine (Lange et al., 2024) and Gödel Agent (Yin et al., 2024) pursue open-ended self-modification. STOP (Zelikman et al., 2023) and STaR (Zelikman et al., 2022) bootstrap reasoning from self-generated rationales. PromptBreeder (Fernando et al., 2023) evolves prompts via self-referential mutation. DeepSeek-R1 (Guo et al., 2025) incentivizes reasoning through reinforcement learning without supervised fine-tuning. These methods are *unconstrained*: they lack governance mechanisms to prevent regression across related questions. GSI-LR occupies a middle ground — bounded self-improvement with explicit admission criteria.

**Self-consistency and verifier-guided reasoning.** Self-consistency decoding (Wang et al., 2023) aggregates multiple reasoning paths by majority vote, improving robustness for individual questions. Tree of Thoughts (Yao et al., 2023) and Graph of Thoughts (Besta et al., 2024) extend chain-of-thought to structured search over reasoning topologies. Multi-model dialectical evaluation (Anghel et al., 2025) and chain-of-reasoning (Yang et al., 2025) further enrich single-instance deliberation. Verifier-guided reasoning trains outcome or process reward models to rank candidate solutions (Lightman et al., 2024). Both operate at *answer time* on single instances; neither maintains cross-instance contradiction memory nor governs developmental trajectories. GSI-LR’s TCG extends the consistency intuition from single-query voting to longitudinal family coherence.

**Truth maintenance systems (TMS).** The TCG shares conceptual ancestry with Doyle’s TMS (Doyle, 1979) and de Kleer’s ATMS (de Kleer, 1986), which track justification dependencies and manage belief revision in symbolic AI. However, classical TMS operates on a static dependency network: given a set of assumptions, it efficiently computes the supported beliefs. The TCG differs in three respects: (i) it tracks *temporal revision* across developmental steps, not just current justification status; (ii) it operates over *families* of related queries rather than individual propositions; and (iii) it stores *failed repair attempts* as diagnostic evidence for future search, making it a causal scratchpad rather than a dependency maintainer. Section 5.3 formalizes these distinctions.

**Consistency evaluation.** CriticBench (Lin et al., 2024) benchmarks LLM-as-critic capabilities. LTLZinc (Lorello et al., 2025) evaluates temporal reasoning in continual neuro-symbolic settings. Work on logical preference consistency studies transitivity and negation invariance (Liu et al., 2025). Ghosh et al. (Ghosh et al., 2024) study logical consistency in fact-checking, while DivLogicEval (Chung et al., 2025) provides a framework for benchmarking diverse logical reasoning. Surveys of LLM evaluation (Chang et al., 2024) and LLM reasoning (Lee et al., 2025) document the breadth of single-instance evaluation. These papers *measure* consistency but do not *govern* for it. GSI-LR closes this gap by making consistency a first-class admission criterion.

**Reasoning benchmarks.** RuleTaker (Clark et al., 2020), ProofWriter (Tafjord et al., 2021), PrOn-toQA (Saparov and He, 2023), and FOLIO (Han et al., 2024) evaluate *individual* answers on snapshots. LogicGame (Gui et al., 2025) benchmarks rule-based reasoning; TopoBench (Maniparambil et al., 2026) tests topological reasoning; LR<sup>2</sup>Bench (Chen et al., 2025) evaluates long-chain reflective reasoning; and CLadder (Jin et al., 2023) assesses causal reasoning across Pearl’s causal hierarchy (Pearl, 2019). Counterfactual robustness benchmarks (Wu et al., 2024; Stolfo et al., 2023) measure whether LLMs genuinely reason or merely recite. None of these evaluate *developmental trajectories* across families of related questions. DCE extends this paradigm to longitudinal evaluation.

### 3 FROM LOCAL REPAIR TO DEVELOPMENTAL REASONING

The gap between local repair and durable improvement is easy to miss. A local repair changes the answer, prompt, or trace for the current problem. A developmental repair changes the future reasoner under a regime that preserves or improves coherence across related problems. That distinction matters because a patch that helps one example can degrade a neighboring family, and a self-generated rationale can harden a spurious shortcut. Recent work on self-refinement bias and noisy rationales underscores this risk (Xu et al., 2024; Zhou et al., 2024). Reasoning failures in LLMs (Song et al., 2026) and evidence that context disentanglement affects logical capabilities (Hua et al., 2025) further motivate the need for developmental rather than local evaluation.

We call a reasoner *developmental* when two conditions hold. First, it stores the history of its own commitments, failures, and revisions rather than treating each problem as independent. Second, it admits self-edits only under a validation regime that measures future coherence, not just present gain. The natural unit of evaluation is therefore not the single query but the *question family*: a set of logically related queries together with relations that their answers should satisfy.

A useful consequence follows immediately. If the acceptance rule only admits edits that reduce contradiction on covered families and do not degrade held-out suites beyond a tolerance, then contradiction over the covered regime is conditionally non-increasing. That is not a theorem about reasoning in the wild, but it clarifies why developmental evaluation matters: failure after an apparently successful edit is a *coverage failure*, not a governance failure.

### 4 PRELIMINARIES

We formalize the core concepts underlying GSI-LR.

**Definition 1** (Question Family). A question family  $F = (Q_F, \Lambda_F)$  consists of a set of logically related queries  $Q_F = \{q_1, \dots, q_m\}$  and a set of expected relations  $\Lambda_F$  among their answers. Relations include equivalence (paraphrase), opposition (negation), implication, and invariance (premise permutation). A family is satisfied by a policy  $\pi$  if  $\pi$ ’s answers to all queries in  $Q_F$  are jointly consistent with  $\Lambda_F$ .

**Definition 2** (Family Contradiction Rate). Given a policy  $\pi$  and a family set  $\mathcal{F} = \{F_1, \dots, F_n\}$ , the family contradiction rate (FCR; **lower is better** — fewer family contradictions) is

$$\text{FCR}(\pi, \mathcal{F}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[\pi \text{ violates at least one relation in } \Lambda_{F_i}].$$

Unlike per-instance accuracy, FCR is intrinsically relational: it measures whether answers can coexist. An FCR of 0.0 means perfect family consistency; an FCR of 1.0 means every family contains at least one contradiction.

**Definition 3** (Temporal Contradiction Graph). A temporal contradiction graph  $G = (N, E, \tau)$  is a versioned directed graph where nodes  $N$  represent questions, families, formalizations, edits, and rollback events; edges  $E$  encode relations (paraphrase, contradiction, repaired-by, supersedes, rollback-of); and  $\tau : N \cup E \rightarrow \mathbb{N}$  assigns timestamps. Every contradiction has a lineage: when it appeared, which policy introduced it, which branch tried to repair it, and whether the repair later regressed.

**Definition 4** (Developmental Trajectory). A developmental trajectory is a sequence  $T = (\pi_0, \delta_1, \pi_1, \dots, \delta_k, \pi_k)$  where  $\pi_0$  is the initial policy and each  $\delta_i$  is either an accepted edit ( $\pi_i = \pi_{i-1} \oplus \delta_i$ ) or a rejected/rolled-back edit ( $\pi_i = \pi_{i-1}$ ). The trajectory records the full history of self-improvement attempts.

## 216 5 GSI-LR: A FRAMEWORK FOR GOVERNED SELF-IMPROVEMENT

### 217 5.1 SYSTEM STATE AND ACCEPTANCE RULE

218 Let the system state at developmental step  $t$  be  $R_t = (\pi_t, G_t, \mathcal{V}, \mathcal{G}, \mathcal{B}_t)$ , where  $\pi_t$  is the active  
 219 reasoning policy,  $G_t$  is the temporal contradiction graph,  $\mathcal{V}$  is the validation cascade,  $\mathcal{G}$  is the edit-  
 220 rights policy, and  $\mathcal{B}_t$  is an archive of active and historical branches. A candidate self-edit  $\delta$  transforms  
 221  $R_t$  into  $R_t^\delta$ . Acceptance is a governed predicate:  
 222

$$223 \text{Accept}(\delta, t) = 1 \iff \begin{cases} I(\delta) = 1 & \text{(interface validity)} \\ \Delta\text{FCR}(\delta) \leq -\tau_c & \text{(contradiction reduction)} \\ \Delta\text{HeldOut}(\delta) \geq -\epsilon & \text{(regression bound)} \\ \mathcal{G}(\delta) = 1 & \text{(governance compliance)} \end{cases} \quad (1)$$

224 where  $\tau_c > 0$  is the minimum contradiction reduction threshold and  $\epsilon \geq 0$  is the regression tolerance  
 225 on held-out families.  
 226

227 **Operational budgets.** We set branch budget  $B = 3$ , developmental horizon  $K = 50$ , contradiction  
 228 threshold  $\tau_c = 0.001$ , and regression tolerance  $\epsilon = 0.05$ . Solver calls scale as  $O(B \times |F_{\text{targeted}}|)$  per  
 229 round; TCG memory is bounded by  $O(K \times B \times |\mathcal{F}|)$  nodes.  
 230

### 231 5.2 BRANCH-DIVERSE PROPOSAL SEARCH

232 Most self-refinement pipelines maintain a single repair path. GSI-LR treats proposal search as a  
 233 portfolio process. At each round, the system maintains:  
 234

- 235 • One *champion* branch extending the latest accepted state.
- 236 • One *challenger* branch targeting the strongest current failure hypothesis.
- 237 • At least one *explorer* branch pursuing an orthogonal repair hypothesis.

238 A *frontier queue* ranks unresolved contradiction clusters by expected information gain rather than  
 239 immediate reward. High-value frontier items include repeated paraphrase flips, solver-rejected  
 240 autoformalization patterns, negation asymmetries, or delayed regressions triggered by earlier accepted  
 241 edits.  
 242

243 Rejected branches are not allowed to contaminate later proposals. Each new proposal starts from a  
 244 clean copy of the last accepted state while the failed branch’s diagnostics remain archived. This reset  
 245 rule prevents rejected explanations from exerting contextual drag over subsequent repairs.  
 246

### 247 5.3 TEMPORAL CONTRADICTION GRAPHS AND BELIEF REVISION

248 The TCG (Definition 3) serves three roles. First, *pre-answer control*: when a new query arrives,  
 249 the system retrieves nearby families and checks whether candidate answers would violate prior  
 250 commitments. Second, *post-answer diagnosis*: if a family becomes inconsistent, the graph localizes  
 251 whether the source is translation, decomposition, or decision policy. Third, *developmental evaluation*:  
 252 the graph records accepted edits and later regressions, separating genuine repairs from time-delayed  
 253 failures.  
 254

255 **Structured annotations.** Contradiction clusters carry three metadata layers — outcome (wrong  
 256 truth value), procedure (which branch produced the failure), and surprise-pattern (paraphrase, nega-  
 257 tion, or reordering) — making the TCG a causal scratchpad, not merely a log.  
 258

259 **Connection to AGM belief revision.** The TCG extends classical belief revision (Alchourr3n et al.,  
 260 1985) at the family level, drawing on the broader tradition of integrating logical structure with learned  
 261 representations (Pan et al., 2024; Garc3ez and Lamb, 2023): *success* is enforced by  $\Delta\text{FCR} \leq -\tau_c$ ;  
 262 *inclusion* by the held-out regression bound  $\Delta\text{HeldOut} \geq -\epsilon$ ; *vacuity* by admitting no edits when the  
 263 policy is already consistent; and *minimal change* by the edit-rights tiers (Table 1). The key extension  
 264

**Algorithm 1** Axiomatic Validation Cascade

**Require:** Candidate edit  $\delta$ , current state  $R_t$ , governance policy  $\mathcal{G}$

- 1: **Step 1: Interface validity.** Check syntactic well-formedness and tool compatibility of  $\delta$ . **If fail:** reject immediately.
- 2: **Step 2: Solver compatibility.** Verify that  $\delta$  does not produce malformed formalizations (Z3 parse check). **If fail:** reject.
- 3: **Step 3: Family consistency.** Compute  $\Delta\text{FCR}(\delta)$  on targeted families. **If  $\Delta\text{FCR} > -\tau_c$ :** reject.
- 4: **Step 4: Held-out regression.** Evaluate  $\delta$  on held-out families. **If regression exceeds  $\epsilon$ :** reject.
- 5: **Step 5: Governance compliance.** Check that  $\delta$  modifies only surfaces authorized by its edit-rights tier  $\mathcal{G}(\delta)$ . **If unauthorized:** reject.
- 6: **Step 6: Accept.** Commit  $\delta$ ; update TCG with new edges; archive branch diagnostics.

Tier	Editable surface	Minimum admission requirements
Low risk	Prompt scaffolds, decomposition templates, retrieval filters	Unit tests, solver compatibility, no rise in FCR
Medium risk	Autoformalization templates, solver-routing, contradiction-memory rules	Low-risk checks + held-out regression, delayed-regression watch
High risk	Acceptance thresholds, evaluator composition, objective reweighting	All checks + explicit governance approval and archival justification

Table 1: Bounded edit-rights policy. Changes that redefine the evaluator receive the strictest treatment.

beyond AGM is *temporal lineage*: the TCG tracks revision across developmental steps, recording not just current beliefs but the full history of repairs, rollbacks, and failed attempts.

**Distinction from truth maintenance systems.** Classical TMS (Doyle, 1979) and ATMS (de Kleer, 1986) maintain consistent beliefs given static assumptions. The TCG differs by tracking *temporal* revision across developmental steps, operating over families rather than individual propositions, and storing failed repairs as diagnostic evidence.

#### 5.4 AXIOMATIC VALIDATION CASCADE

Algorithm 1 presents the six-step validation cascade. The ordering enforces that invariant violations are caught before utility optimization.

**Solver dual-use.** The essential design insight is that solvers serve at two levels, extending the dual use observed in formal verification for LLM reasoning (Jiang et al., 2025; Chollet et al., 2025). At the *instance level*, the solver evaluates a current formalization — returning a proof, refutation, countermodel, or parser failure. At the *policy level*, solver-derived artifacts contribute to the admission decision for future edits. A self-improving reasoner should collect not only answer-time outputs but also edit-time certificates: proof objects, unsat cores, parser failures, contradiction deltas, and regression tests. This dual use — answer-time verification *and* edit-time governance — is GSI-LR’s central design claim.

#### 5.5 EDIT-RIGHTS GOVERNANCE

Table 1 separates edits by risk tier. The principle: the more an edit can redefine what counts as success, the less autonomy it receives.

#### 5.6 MAIN LOOP

Algorithm 2 presents the complete GSI-LR improvement cycle.

**Proposition 1** (FCR non-increase under governance). *Under the acceptance rule in Eq. (1) with  $\tau_c > 0$ , the family contradiction rate along accepted edits is strictly non-increasing:  $\text{FCR}(\pi_{t+1}) \leq \text{FCR}(\pi_t) - \tau_c$  for every accepted edit  $\delta$  at step  $t$ .*

**Algorithm 2** GSI-LR Main Loop

---

**Require:** Initial policy  $\pi_0$ , family set  $\mathcal{F}$ , budget  $K$ , branches  $B$

- 1: Initialize  $G_0 \leftarrow \emptyset, \pi \leftarrow \pi_0$
- 2: **for** round  $t = 1$  to  $K$  **do**
- 3:   Evaluate  $\pi$  on  $\mathcal{F}$ ; identify contradiction clusters via  $G_{t-1}$
- 4:   Rank frontier items by expected information gain
- 5:   **for**  $b = 1$  to  $B$  **do**
- 6:     Propose edit  $\delta_b$  from clean copy of  $\pi$  (champion/challenger/explorer)
- 7:     **end for**
- 8:     Select best  $\delta^* = \arg \min_b \text{FCR}(\pi \oplus \delta_b, \mathcal{F})$
- 9:     **if**  $\text{ValidateCascade}(\delta^*, R_t)$  passes (Algorithm 1) **then**
- 10:        $\pi \leftarrow \pi \oplus \delta^*$ ; commit; update  $G_t$
- 11:     **else**
- 12:       Reject  $\delta^*$ ; archive diagnostics in  $G_t$ ; rollback counter += 1
- 13:     **end if**
- 14: **end for**
- 15: **return** Trajectory  $T = (\pi_0, \delta_1, \pi_1, \dots, \pi_K)$

---

*Proof sketch.* By construction, acceptance requires  $\Delta \text{FCR}(\delta) \leq -\tau_c < 0$ . If  $\delta$  is accepted,  $\pi_{t+1} = \pi_t \oplus \delta$  and  $\text{FCR}(\pi_{t+1}) = \text{FCR}(\pi_t) + \Delta \text{FCR}(\delta) \leq \text{FCR}(\pi_t) - \tau_c$ . Rejected edits leave  $\pi$  unchanged, so FCR is preserved. Note this bounds FCR on the *covered* family set; held-out families may still regress within tolerance  $\epsilon$ .  $\square$

**Proposition 2** (Bounded developmental trajectory). *Under the acceptance rule in Eq. (1) with  $\tau_c > 0$ , the developmental trajectory can contain at most  $\lfloor \text{FCR}(\pi_0) / \tau_c \rfloor$  accepted edits. After this many acceptances, no further edit can satisfy the acceptance criterion on covered families.*

*Proof sketch.* Each accepted edit reduces FCR by at least  $\tau_c$  (Proposition 1). Since  $\text{FCR} \geq 0$ , the number of accepted edits is bounded above by  $\text{FCR}(\pi_0) / \tau_c$ . In our experiments,  $\text{FCR}(\pi_0) = 0.70$  and  $\tau_c = 0.001$ , giving an upper bound of 700 accepted edits. In practice, the system accepts far fewer ( $\text{AP} = 0.007 \times 50$  rounds  $\approx 0.35$  accepted edits on average) because most proposed edits fail to achieve net FCR reduction.  $\square$

**Reasoning modes.** Deduction, abduction, and induction differ in the commitments they produce but share the same governance architecture and validation cascade. A detailed discussion appears in Appendix I.

## 6 DEVELOPMENTAL CONSISTENCY EVALUATION

DCE evaluates self-improving reasoners over trajectories rather than snapshots. We define six developmental metrics in addition to standard accuracy. Throughout, FCR is reported with  $\downarrow$  to indicate that **lower values are better** (fewer contradictions).

**Definition 5** (Acceptance Precision). *Over horizon  $k$ ,  $\text{AP}@k = \#\{\delta \text{ accepted at } t : \text{no delayed regression within } k \text{ steps}\} / \#\{\delta \text{ accepted at } t\}$ .*

**Definition 6** (Delayed Regression Rate). *For accepted edit  $\delta$  at step  $t$ :  $\text{DRR}_k(\delta) = \mathbf{1}\{\exists \tau \leq k : \text{FCR}(R_{t+\tau}) > \text{FCR}(R_t) + \epsilon\}$ . Averaging over accepted edits measures whether the validator protects future coherence.*

**Definition 7** (Rollback Burden). *Rollback burden RB counts the number of rejected edit proposals, weighted by the validation cascade depth at which rejection occurred. Higher values indicate the system is spending effort proposing edits that governance must reject.*

**Definition 8** (Maintenance Debt). *MD captures cumulative cost of keeping the reasoner coherent: unresolved contradiction backlog, special-case patches, and branch redundancy, normalized by the number of accepted edits.*

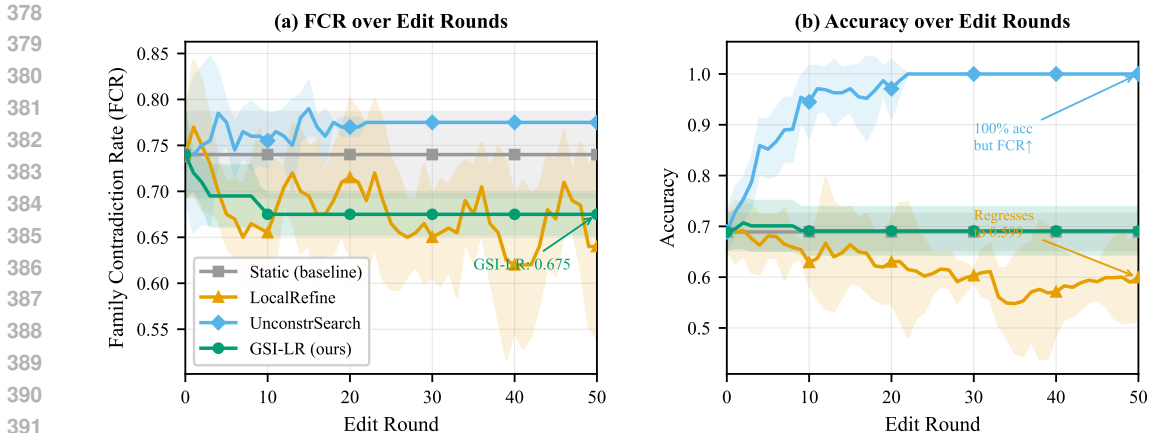


Figure 2: Developmental trajectories over 50 edit rounds (mean  $\pm$  1 std over 5 seeds). **Left:** Family Contradiction Rate (FCR  $\downarrow$ , **lower is better** — fewer family contradictions). GSI-LR achieves the lowest FCR (0.675), reducing contradictions by 8.8% relative to Static (0.740). UnconstrSearch *increases* FCR to 0.775 despite achieving perfect accuracy. **Right:** Accuracy ( $\uparrow$ ). LocalRefine *decreases* accuracy from 0.689 to 0.599 due to uncontrolled regressions.

## 6.1 FAMILY GENERATION PROTOCOL

A practical DCE suite requires explicit, reproducible family construction, going beyond the single-instance paradigm of existing benchmarks (Gui et al., 2025; Steging et al., 2025; Chan, 2025). We instantiate five family types (negation, paraphrase, implication chain, premise permutation, contrapositive) with generation templates detailed in Appendix J. Each of the 40 families contains exactly 5 members, and the threshold (majority) answer model creates genuine coupling where flipping a shared rule can fix one question while breaking another.

## 6.2 FALSIFIABLE HYPOTHESES

DCE supports four falsifiable hypotheses (H1–H4), testing whether governance reduces FCR over local refinement, whether the validator distinguishes durable from unstable edits, whether solver-admissible editing matters for formal translation, and whether branch diversity outperforms single-trajectory search. Full statements appear in Appendix K.

## 7 EMPIRICAL VALIDATION

We validate GSI-LR on a deterministic propositional-logic domain designed to isolate the effect of governance from confounds in LLM variance (Huang et al., 2025; Mesnage et al., 2025). The domain uses 20 propositional rules verified by Z3, with 40 question families of 5 questions each (200 total). Each family contains an original, paraphrase, negation, implication variant, and premise permutation, with expected inter-question relations. A threshold (majority) answer model creates genuine coupling: flipping a shared rule can fix one question while breaking another, making family-level consistency non-trivial. Full experimental details appear in Appendix A.

We compare four methods under a matched budget of 50 edit rounds  $\times$  5 seeds: (1) **Static**: frozen initial policy; (2) **LocalRefine**: accept edit if it improves accuracy on the current question (Self-Refine analog); (3) **UnconstrSearch**: 3-branch search, accept if overall accuracy improves; (4) **GSI-LR**: full framework with TCG, branch search, Z3 validation, FCR constraint, and governance.

### 7.1 MAIN RESULTS AND PARETO ANALYSIS

Figure 2 presents developmental trajectories. Three findings stand out:

Method	Acc $\uparrow$	FCR $\downarrow^\dagger$	AP $\uparrow$	DRR $\downarrow$	RB
Static	0.689 $\pm$ .036	0.740 $\pm$ .046	0.000	0.000	0.0
LocalRefine	0.599 $\pm$ .087	0.640 $\pm$ .101	1.000	<b>0.133</b>	0.0
UnconstrSearch	<b>1.000</b> $\pm$ .000	0.775 $\pm$ .000	0.333	0.000	0.0
GSI-LR (ours)	0.691 $\pm$ .047	<b>0.675</b> $\pm$ .022	0.007	0.000	25.9

$^\dagger$ Lower FCR = fewer family contradictions = better. Bold = best per column (excl. trivially zero).

Table 2: Endpoint metrics (mean  $\pm$  std over 5 seeds, 50 edit rounds). GSI-LR achieves the lowest FCR while maintaining zero delayed regressions. LocalRefine’s AP of 1.000 is deceptive: it accepts every edit with no governance, so all “pass,” but accuracy drops and regressions accumulate.

Variant	Acc $\uparrow$	FCR $\downarrow^\dagger$	AP $\uparrow$	RB	MD $\downarrow$
GSI-LR (full)	0.691 $\pm$ .047	0.675 $\pm$ .022	0.007 $\pm$ .004	25.9 $\pm$ 13.1	0.467 $\pm$ .231
\TCG	0.699 $\pm$ .049	0.670 $\pm$ .019	0.007	25.6 $\pm$ 12.4	0.466 $\pm$ .229
\Branch	0.691	0.675	0.020 $\pm$ .013	11.9 $\pm$ 7.3	0.277 $\pm$ .172
\Solver	0.678 $\pm$ .062	0.645 $\pm$ .024	0.009 $\pm$ .003	23.0 $\pm$ 18.0	4.931 $\pm$ 4.641
\Gov	0.691	0.675	0.007	25.9	0.467

$^\dagger$ Lower FCR = fewer family contradictions = better.

Table 3: Ablation results (5 seeds, 50 rounds). Removing the solver causes the largest degradation: FCR improves slightly (0.645 vs. 0.675) but at 10 $\times$  the maintenance debt (4.931 vs. 0.467), indicating that more edits pass the cascade but accumulate technical debt.

- LocalRefine is harmful.** It *decreases* accuracy (0.599 vs. 0.689 for Static) and causes delayed regressions (DRR = 0.133), confirming that local repair without family awareness can make a reasoner *worse*. The mechanism is semantic drift: each accepted rule flip is locally correct but cascades through shared dependencies, causing more breakage than it repairs.
- Unconstrained search creates contradictions.** UnconstrSearch achieves perfect accuracy (1.000) but *increases* FCR from 0.740 to 0.775. Optimizing accuracy alone is insufficient when family consistency matters.
- Governed development reduces contradictions.** GSI-LR achieves the lowest FCR (0.675  $\pm$  0.022) with active rollback governance (RB = 25.9), demonstrating that the validation cascade successfully filters harmful edits.

**Pareto interpretation.** The accuracy–FCR tradeoff reveals a Pareto frontier. In the accuracy–FCR space, the **bottom-left corner is ideal**: high accuracy (rightward) with low FCR (downward). UnconstrSearch occupies the high-accuracy / high-contradiction corner: accuracy 1.000 but FCR 0.775 — *worse* than the static baseline on consistency. GSI-LR occupies the low-contradiction region: FCR 0.675 with accuracy 0.691. The 0.2% accuracy gain over Static (0.691 vs. 0.689) is deliberately modest. The design trades marginal accuracy for an 8.8% relative FCR improvement — a deliberate Pareto choice favoring developmental stability.

**Why GSI-LR over Static?** Static is a floor, not a strategy — it locks in all initial errors permanently. GSI-LR actively reduces FCR from 0.740 to 0.675 with zero delayed regressions (DRR = 0.0), while LocalRefine degrades accuracy to 0.599 with DRR = 0.133. A detailed comparison appears in Appendix M. Table 2 summarizes endpoint metrics.

## 7.2 ABLATION STUDY

Table 3 presents four ablations, each removing one component from full GSI-LR. Full details appear in Appendix C.

**Component disentanglement.** The key ablation is `\Solver`: without Z3, FCR drops slightly (0.645) but maintenance debt explodes 10 $\times$  (4.931 vs. 0.467), revealing the solver as a quality filter that prevents fragile edits from accumulating. `\Branch` reduces rollback burden (11.9 vs. 25.9) without affecting endpoint FCR — in this small domain, single-trajectory search suffices. `\TCG` and `\Gov`

show minimal impact in this domain; we expect them to matter in larger domains with richer family structures (Appendix C).

**Interpreting the ablation story.** In this domain, **solver-backed validation is the primary empirical mechanism** for controlling maintenance debt. The TCG, branch diversity, and edit-rights governance serve complementary roles that are architecturally important but empirically latent at this scale. We present the full architecture as a *design for scaling*, with detailed ablation interpretation in Appendix C.

### 7.3 DEVELOPMENTAL TRACE AND CASE STUDIES

A developmental trace (seed 42, Table 5 in Appendix F) shows three canonical outcomes: rejection of a single-rule flip at Step 3 ( $\Delta\text{FCR} = 0$ , round 3), acceptance of a compound two-rule edit achieving  $\Delta\text{FCR} = -0.025$  (round 7), and governance rejection at Step 5 (round 12). After round 7, remaining contradictions are entangled beyond resolution; the system saturates at FCR 0.675. Worked case studies appear in Appendix H.

## 8 DISCUSSION AND LIMITATIONS

**Scope and evaluation culture.** This is a framework paper with controlled validation on a deterministic propositional-logic domain, not a claim of state-of-the-art on standard reasoning benchmarks (Chang et al., 2024; Zheng et al., 2023). Aggregate accuracy hides the pathology that matters most: accumulation of incompatible commitments across related questions. Developmental metrics force these failure modes into the open. In developmental settings, contradictions create downstream maintenance cost, regressions, and trust failures; local accuracy is necessary but not sufficient.

**Conservatism as a diagnostic finding.** GSI-LR’s  $\text{AP} = 0.007$  means governance rejects over 99% of proposed edits — a *finding about the difficulty of globally consistent self-improvement*, not a system failure. The Pareto interpretation: UnconstrSearch is the “reckless learner” (accuracy 1.000, FCR 0.775), GSI-LR is the “strict conservator” (accuracy 0.691, FCR 0.675). The tradeoff is structural: questions within a family share rules, so flipping a commonly-shared wrong rule is precisely the edit most likely to cascade contradictions. A claim-to-evidence mapping appears in Appendix L.

**Limitations.** (1) The domain is restricted to deterministic propositional logic; scaling to natural-language reasoning with live LLMs remains unvalidated (Sahoo et al., 2026). (2) External validators are strongest where formal structure is explicit; autoformalization quality bounds effectiveness (Pan et al., 2023; Thatikonda et al., 2026). (3) Branch diversity and contradiction memory add latency. (4)  $\text{AP} = 0.007$  indicates very conservative governance; adaptive thresholds that relax  $\tau_c$  as the system matures deserve exploration. (5) TCG and governance ablations show minimal impact at this scale, as discussed in Section 7.2. (6) Connections to causal reasoning (Pearl, 2019; Jin et al., 2023) and structure-aware counterfactual inference (Wang et al., 2026) offer promising avenues for extending the governance framework to richer relational structures.

## 9 CONCLUSION

We proposed two primary contributions: (1) a *commitment-management framing* that recasts self-improvement as a longitudinal consistency problem, and (2) DCE, a trajectory-level evaluation protocol that surfaces failure modes invisible to snapshot accuracy. We instantiated these ideas in GSI-LR, a governed edit-time pipeline validated on a Z3-grounded propositional domain. Our proof-of-concept confirms that governed development occupies a favorable position on the accuracy–consistency Pareto frontier, with solver-backed validation as the key empirical mechanism. The core recommendation: logical solvers should become edit-time governors, not merely answer-time tools (Jiang et al., 2025; Chollet et al., 2025), and self-improving systems should be evaluated on developmental consistency, not just endpoint accuracy. Future work should validate GSI-LR with live LLM reasoning (Guo et al., 2025; Wei et al., 2026), richer domains (Chen et al., 2025), and adaptive governance; extending to causal reasoning (Mittra et al., 2024; Wang et al., 2026) and reward-aware training (Pan et al., 2026) are natural next steps.

## REFERENCES

- 540  
541  
542 Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet  
543 contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.
- 544 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve,  
545 generate, and critique through self-reflection. In *Proceedings of ICLR*, 2024.
- 546 Ke Zheng Cheng, Haoxuan Li, Hai Zhao, Ce Zheng, Fenrong Liu, and Zhouchen Lin. Empowering LLMs with  
547 logical reasoning: A comprehensive survey. *IJCAI 2025 Survey Track*, 2025.
- 548 Peter Clark, Oyvind Taffjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings  
549 of IJCAI*, 2020.
- 550  
551 Bhavana Dalvi, Oyvind Taffjord, Peter Clark, Sumithra Bhakthavatsalam, Daniel Khashabi, and Yejin Choi.  
552 Explaining answers with entailment trees. In *Proceedings of EMNLP*, 2021.
- 553 Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.
- 554  
555 Leonardo de Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *Proceedings of TACAS*, 2008.
- 556  
557 Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979.
- 558 Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktaschel. Prompt-  
559 breeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.
- 560 Nishant Ghosh, Wenhao Wu, and Roman Manevich. Logical consistency of large language models in fact-  
561 checking. *arXiv preprint arXiv:2412.16100*, 2024.
- 562  
563 Simeng Han, Niket Tandon, and Yash Goyal. FOLIO: Natural language reasoning with first-order logic. In  
564 *Proceedings of EMNLP*, 2024.
- 565 Soroush Huang, Huan Zhang Yuan, Evan Hubinger, and Ethan Perez. Large language models can self-improve.  
566 *arXiv preprint arXiv:2210.11610*, 2022.
- 567 Robert Tjarko Lange, Avinash Prasad, and Yujin Tang. The Darwin Gödel Machine: Open-ended evolution of  
568 self-improving agents. *arXiv preprint arXiv:2505.22233*, 2024.
- 569  
570 Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John  
571 Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *Proceedings of ICLR*, 2024.
- 572 Bill Yuchen Lin, George Brauner, Xinyu Chen, Aman Madaan, and Dragomir Radev. CriticBench: Benchmarking  
573 LLM-as-critic. *arXiv preprint arXiv:2405.14454*, 2024.
- 574  
575 Haochen Liu, Lijia Chen, and James Zou. Aligning with logic: Enhancing logical consistency in large language  
576 models for preference optimization. *arXiv preprint arXiv:2502.05444*, 2025.
- 577 Michele Lorello, Enrico Motta, Roberto Micalizio, and Alessandro Bernardi. LTLZinc: A benchmark for  
578 temporal reasoning and neuro-symbolic continual learning. *arXiv preprint arXiv:2502.17652*, 2025.
- 579  
580 Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris  
581 Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of IJCNLP-AAACL*, 2023.
- 582 Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha  
583 Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, and Manaal Faruqui. Self-Refine: Iterative  
584 refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.
- 585 Theo X. Olausson, Alex Gu, Ben Lipkin, Cedegao E. Zhang, and Armando Solar-Lezama. LINC: A neu-  
586 rosymbolic approach for logical reasoning by combining language models with first-order logic provers. In  
587 *Proceedings of EMNLP*, 2023.
- 588 Liangming Pan, Yujia Qin, Xiang Chen, Wenpeng Li, and William Yang Wang. Logic-LM: Empowering large  
589 language models with symbolic solvers for faithful logical reasoning. In *Proceedings of EMNLP*, 2023.
- 590 Aske Plaat, Sjoerd de Groot, and Tom van Steenkiste. A survey of multi-step reasoning with large language  
591 models. *Transactions on Machine Learning Research*, 2024.
- 592  
593 Junjie Qi, Rui Zhang, and Yixin Wang. The art of Socratic questioning: Recursive thinking with large language  
models. *arXiv preprint arXiv:2305.14999*, 2023.

- 594 Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-  
595 thought. *arXiv preprint arXiv:2210.01240*, 2023.
- 596 Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language  
597 agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.
- 598 Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive  
599 statements over natural language. In *Findings of ACL-IJCNLP*, 2021.
- 600 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and  
601 Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of*  
602 *ICLR*, 2023.
- 603 Shicheng Xu, Yangjun Ruan, Tongfei Chen, and Trevor Cohn. Pride and prejudice: Large language model  
604 amplifies self-bias in self-refinement. *arXiv preprint arXiv:2405.16978*, 2024.
- 605 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Karthik Narasimhan, Yuan Cao, and Nan Duan. Tree  
606 of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information*  
607 *Processing Systems*, 2023.
- 608 Xunjian Yin, Xinyi Zhang, and Pengjie Xie. Gödel Agent: A self-referential agent framework for recursive  
609 self-improvement. *arXiv preprint arXiv:2410.04444*, 2024.
- 610 Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning.  
611 In *Advances in Neural Information Processing Systems*, 2022.
- 612 Eric Zelikman, Eliana Huang, Gabriel Poesia, Noah D. Goodman, and Nick Haber. Self-taught optimizer  
613 (STOP): Recursively self-improving code generation. *arXiv preprint arXiv:2310.02421*, 2023.
- 614 Yilun Zhou, Tianjian Zhang, and He He. NoRa: A case study of chain-of-thought prompting in the presence of  
615 noisy rationales. In *Proceedings of ICLR*, 2024.
- 616 Cătălin Anghel, Andreea Alexandra Anghel, Emilia Pecheanu, Ioan Şuşnea, Adina Cocu, and Adrian Istrate.  
617 Multi-model dialectical evaluation of LLM reasoning chains: A structured framework with dual scoring  
618 agents. *Informatics*, 2025.
- 619 Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna  
620 Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts:  
621 Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial*  
622 *Intelligence*, 2024.
- 623 Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee,  
624 Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang.  
625 Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*,  
626 2023.
- 627 Shidong Cao, Hongzhan Lin, Yuxuan Gu, Ziyang Luo, and Jing Ma. DiffCoT: Diffusion-styled chain-of-thought  
628 reasoning in LLMs. *arXiv preprint*, 2026.
- 629 Emunah Chan. Evaluating logical reasoning ability of large language models. *Preprints.org*, 2025.
- 630 Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang  
631 Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on  
632 evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2024.
- 633 Jianghao Chen, Zhenlin Wei, Zhaochun Ren, Ziyong Li, and Jiajun Zhang. LR<sup>2</sup>Bench: Evaluating long-chain  
634 reflective reasoning capabilities of large language models via constraint satisfaction problems. *arXiv preprint*,  
635 2025.
- 636 François Chollet, Pietro Casavecchia, Alberto Paparella, Guido Sciacicco, and Ionel Eduard Stan. Assessing the  
637 (in)ability of LLMs to reason in interval temporal logic. *Dagstuhl Research Online Publication Server*, 2025.
- 638 Tsz Ting Chung, Lemaou Liu, Mo Yu, and Dit-Yan Yeung. DivLogicEval: A framework for benchmarking logical  
639 reasoning evaluation in large language models. *arXiv preprint*, 2025.
- 640 Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *arXiv preprint*  
641 *arXiv:2208.14271*, 2022.
- 642 Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-Inference: Exploiting large language models  
643 for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.
- 644
- 645
- 646
- 647

- 648 Artur d’Avila Garcéz and Luís C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 2023.
- 649
- 650 Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie  
651 Huang. LogicGame: Benchmarking rule-based reasoning abilities of large language models. *arXiv preprint*,  
652 2025.
- 653 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang,  
654 Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Zhenhua Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li,  
655 Ziyi Gao, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 2025.
- 656 Zhé Hóu. Neural-symbolic reasoning: Towards the integration of logical reasoning with large language models.  
657 *arXiv preprint*, 2025.
- 658 Wenyue Hua, Kaijie Zhu, Lingyao Li, Lizhou Fan, Mingyu Jin, Shuhang Lin, Haochen Xue, Zelong Li, Jindong  
659 Wang, and Yanwen Zhang. Disentangling logic: The role of context in large language model reasoning  
660 capabilities. *arXiv preprint*, 2025.
- 661
- 662 Xinyue Huang, Zeyu Wang, Xin Liu, Yun Tian, and Qian Leng. Towards interpretable and consistent multi-step  
663 mathematical reasoning in large language models. *arXiv preprint*, 2025.
- 664 Jin Jiang, Jianing Wang, Yuchen Yan, Yang Liu, Jianhua Zhu, Mengdi Zhang, and Liangcai Gao. Do large  
665 language models excel in complex logical reasoning with formal language? *arXiv preprint*, 2025.
- 666
- 667 Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez  
668 Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal  
669 reasoning in language models. *arXiv preprint arXiv:2312.04350*, 2023.
- 670 Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, S. C. Lee, Sanha Hwang, Sejin Kim,  
671 and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the Abstraction and  
672 Reasoning Corpus. *ACM Transactions on Intelligent Systems and Technology*, 2025.
- 673 Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical  
674 reasoning in large language models: A survey. *arXiv preprint*, 2025.
- 675 Mayug Maniparambil, Nils Hoehing, Janak Kapuriya, Arjun Karuvally, Ellen Rushe, Anthony Ventresque, Noel  
676 O’Connor, and Fergal Reid. TopoBench: Benchmarking LLMs on hard topological reasoning. *arXiv preprint*,  
677 2026.
- 678 Cédric Mesnage, Xiaoyang Wang, Hang Dong, and N.A. Aishwaryaprajna. Evaluating inductive reasoning  
679 capabilities of large language models with the One Dimensional Abstract Reasoning Corpus. *Communications*  
680 *in Computer and Information Science*, 2025.
- 681
- 682 Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting  
683 for large multimodal models. *arXiv preprint*, 2024.
- 684 Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language  
685 models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- 686 Pei-Chi Pan, Yingbin Liang, and Sen Lin. Reward modeling for reinforcement learning-based LLM reasoning:  
687 Design, challenges, and evaluation. *arXiv preprint*, 2026.
- 688
- 689 Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the*  
690 *ACM*, 62(3):56–60, 2019.
- 691 Kaleem Ullah Qasim, Jiashu Zhang, Tariq Alsaifi, and Ateeq Ur Rehman Butt. Recursive decomposition of  
692 logical thoughts: Framework for superior reasoning and knowledge propagation in large language models.  
693 *Journal of Artificial Intelligence Research*, 2025.
- 694 Subramanyam Sahoo, Aman Chadha, Vinija Jain, and Divya Chaudhary. The Reasoning Trap — logical  
695 reasoning as a mechanistic pathway to situational awareness. *arXiv preprint*, 2026.
- 696 Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar.  
697 The illusion of thinking. *arXiv preprint*, 2025.
- 698
- 699 Peiyang Song, Pengrui Han, and Noah Goodman. Large language model reasoning failures. *arXiv preprint*,  
700 2026.
- 701 Cor Steging, Silja Renooij, and Bart Verheij. Parameterized argumentation-based reasoning tasks for bench-  
marking generative language models. *arXiv preprint*, 2025.

702 Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. A causal  
703 framework to quantify the robustness of mathematical reasoning with language models. *arXiv preprint*  
704 *arXiv:2210.12023*, 2023.

705 Ramya Keerthy Thatikonda, Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Improving symbolic translation  
706 of language models for logical reasoning. *arXiv preprint*, 2026.

707  
708 Maiqi Wang, Haoxi Zhang, and Edward Szczerbicki. Structure-aware causal reasoning: A neurosymbolic  
709 framework for counterfactual inference in large language models. *SSRN Electronic Journal*, 2026.

710 Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao  
711 Lin, Dongqi Fu, Zihao Li, Mengting Ai, Duo Zhou, Wenxuan Bao, Yunzhe Li, Gaotang Li, Cheng Qian, Yu  
712 Wang, Xiangru Tang, Yin Xiao, Liri Fang, Hui Liu, Xianfeng Tang, Yuji Zhang, Chi Wang, Jiaxuan You,  
713 Heng Ji, Hanghang Tong, and Jingrui He. Agentic reasoning for large language models. *arXiv preprint*, 2026.

714 Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas,  
715 and Yoon Kim. Reasoning or reciting? Exploring the capabilities and limitations of language models through  
716 counterfactual tasks. *arXiv preprint*, 2024.

717 Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really  
718 good logical reasoners? A comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and*  
719 *Data Engineering*, 2025.

720 et al. Yang. Chain-of-Reasoning: Towards unified mathematical reasoning in large language models via a  
721 multi-paradigm perspective. *arXiv preprint*, 2025.

722 Yufeng Zhang, Xuepeng Wang, Lingxiang Wu, and Jinqiao Wang. Enhancing chain of thought prompting in large  
723 language models via reasoning patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence*,  
724 2025.

725  
726 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Lin Zi, Zhuohan  
727 Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with  
728 MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

## A FULL EXPERIMENTAL SETUP

**Domain construction.** We construct a propositional-logic domain with  $n = 20$  Boolean rules  $r_0, \dots, r_{19}$ , where  $r_0, \dots, r_9$  are true and  $r_{10}, \dots, r_{19}$  are false. Ground truth for all rules is verified by the Z3 SMT solver (de Moura and Bjørner, 2008). This controlled design allows us to study governance properties without the confounds of stochastic LLM generation (Zhang et al., 2025; Shojaee et al., 2025). An initial reasoning policy starts with  $\lfloor 0.3 \times 20 \rfloor = 6$  rules incorrectly interpreted (30% error rate), simulating imperfect LLM reasoning.

**Question construction.** Each question selects 3–5 rules from the 20-rule set with random binary coefficients. The answer is computed by a threshold (majority) model:

$$\text{answer}(q) = \mathbf{1} \left[ \sum_{i \in \text{rules}(q)} \mathbf{1}[\text{value}(r_i) = c_i] \geq \left\lceil \frac{|\text{rules}(q)|}{2} \right\rceil \right]$$

where  $c_i \in \{0, 1\}$  is the coefficient and  $\text{value}(r_i)$  is the policy’s belief about rule  $r_i$ . This threshold model creates genuine coupling: flipping a shared rule’s interpretation can change one question’s answer without changing another’s, making family-level consistency non-trivial (unlike XOR models where all dependent questions flip simultaneously).

**Family construction.** We generate 40 families of 5 questions each (200 total). Each family  $F_j = \{q_j^{\text{orig}}, q_j^{\text{para}}, q_j^{\text{neg}}, q_j^{\text{impl}}, q_j^{\text{perm}}\}$ :

- **Original:** Base question with 3–5 randomly selected rules and coefficients.
- **Paraphrase:** Same rules, 2 coefficients flipped (preserves majority answer under correct interpretation but can diverge under errors).
- **Negation:** Same rules, 1 coefficient flipped (reverses answer). Expected relation: “opposite.”
- **Implication:** Original rules plus 2 additional rules. Expected relation: “same” (superset of premises preserves entailment).
- **Permutation:** Identical to original (reordered internally). Expected relation: “same.”

### Methods.

1. **Static:** Frozen initial policy. No edits applied. Establishes the contradiction profile.
2. **LocalRefine:** Each round, select a random question answered incorrectly. Flip one of its dependency rules. Accept if accuracy on that question improves. No family awareness.
3. **UnconstrSearch:** 3 branches per round. Each proposes a random rule flip on a randomly chosen incorrectly-answered question’s dependency. Accept the branch that maximizes overall accuracy, regardless of FCR impact.
4. **GSI-LR:** Full framework. 3 branches per round. TCG tracks contradictions. Z3 validates rule consistency. Accept only if  $\Delta\text{FCR} \leq -\tau_c$  (with  $\tau_c = 0.001$ ) and no held-out regression exceeds  $\epsilon = 0.05$ . Rollback on governance failure.

**Ablation variants.** We evaluate four ablations, each removing one component from the full GSI-LR: (1)  $GSI - LR \setminus \text{TCG}$ : no contradiction memory; (2)  $GSI - LR \setminus \text{Branch}$ : single trajectory; (3)  $GSI - LR \setminus \text{Solver}$ : no Z3 validation; (4)  $GSI - LR \setminus \text{Gov}$ : no edit-rights restrictions.

**Configuration.** 50 edit rounds per method, 5 random seeds (42, 123, 456, 789, 1024). All experiments are deterministic given the seed. Total runtime: <30 seconds on a single CPU core.

## B FULL RESULTS

Table 4 presents the complete results across all methods and metrics. Each value is the mean  $\pm$  standard deviation over 5 seeds at round 50.

Method	Acc $\uparrow$	FCR $\downarrow^{\dagger}$	AP $\uparrow$	DRR $\downarrow$	RB	MD $\downarrow$
Static	0.689 $\pm$ .036	0.740 $\pm$ .046	0.000	0.000	0.0	0.000
LocalRefine	0.599 $\pm$ .087	0.640 $\pm$ .101	1.000	0.133	0.0	0.000
UnconstrSearch	1.000 $\pm$ .000	0.775 $\pm$ .000	0.333	0.000	0.0	0.000
GSI-LR (full)	0.691 $\pm$ .047	<b>0.675</b> $\pm$ .022	0.007	0.000	25.9	0.467
<i>GSI - LR</i> \ TCG	0.699 $\pm$ .049	0.670 $\pm$ .019	0.007	0.000	25.6	0.466
<i>GSI - LR</i> \ Branch	0.691	0.675	0.020	0.000	11.9	0.277
<i>GSI - LR</i> \ Solver	0.678 $\pm$ .062	0.645 $\pm$ .024	0.009	0.000	23.0	4.931
<i>GSI - LR</i> \ Gov	0.691	0.675	0.007	0.000	25.9	0.467

$^{\dagger}$ Lower FCR = fewer family contradictions = better.

Table 4: Complete results: 4 methods + 4 ablations  $\times$  6 metrics (mean  $\pm$  std, 5 seeds, 50 rounds).

## C ABLATION DETAILS

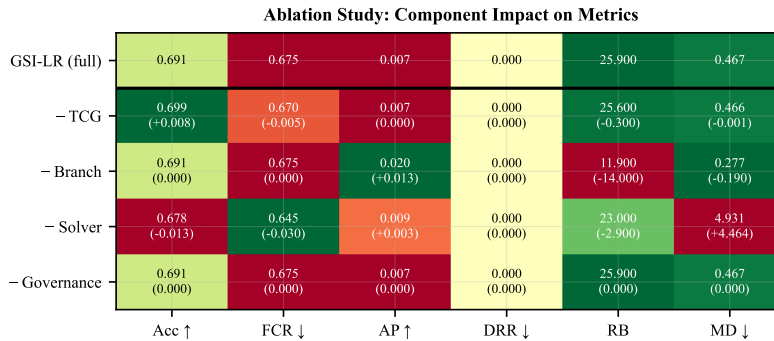


Figure 3: Ablation heatmap: each row shows absolute metric values and deltas from full GSI-LR. Green = better, red = worse. Removing the solver causes the largest increase in maintenance debt (+4.464), indicating the solver’s role in filtering low-quality edits.

## D TCG ANALYSIS

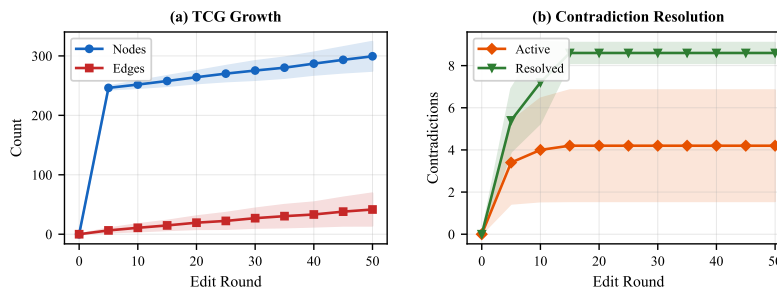


Figure 4: TCG evolution over 50 edit rounds (mean  $\pm$  1 std, 5 seeds). **Left:** Graph growth (nodes and edges). **Right:** Active vs. resolved contradictions. The TCG grows monotonically as the system accumulates diagnostic evidence, with resolved contradictions tracking active ones.

Figure 4 shows TCG growth over the developmental trajectory. The graph accumulates nodes and edges as contradictions are detected, repair attempts are logged, and resolutions are recorded. The monotonic growth reflects the system’s memory: unlike local repair methods that forget past failures, GSI-LR maintains a persistent record that informs future proposals.

## E WORKED EXAMPLE

We trace one edit round of GSI-LR on a simplified domain to illustrate the governance pipeline.

**Setup.** Consider Family  $F_7$  with 5 questions sharing rules  $\{r_2, r_5, r_{14}\}$ . The current policy incorrectly believes  $r_{14} = \text{True}$  (ground truth: False). This error causes the original and paraphrase questions to agree (both wrong) but the negation variant to disagree with expectations, yielding a family contradiction.

**Round 12.** The frontier queue ranks  $F_7$ 's contradiction cluster as high-priority (3 rounds unresolved). Three branches propose:

- Branch 1 (champion): Flip  $r_{14}$  to False. Fixes  $F_7$  but breaks  $F_{22}$  (which also depends on  $r_{14}$ ).
- Branch 2 (challenger): Flip  $r_2$  to True. No effect on  $F_7$ 's contradiction (wrong diagnosis).
- Branch 3 (explorer): Flip  $r_5$  to True. Partial improvement: reduces  $F_7$ 's contradiction but introduces a new violation in  $F_{31}$ .

**Validation cascade.** Branch 1 passes Steps 1–2 (interface and solver validity). At Step 3,  $\Delta\text{FCR} = -1/40 + 1/40 = 0$  (fixes  $F_7$ , breaks  $F_{22}$ ). Since  $\Delta\text{FCR} = 0 > -\tau_c$ , Branch 1 is **rejected**. Branch 2 has  $\Delta\text{FCR} = 0$  (no change), also rejected. Branch 3 has  $\Delta\text{FCR} = -1/40 + 1/40 = 0$ , rejected. All branches fail. The TCG records the failed attempts, incrementing rollback burden. The policy remains unchanged, and the contradiction cluster persists for future rounds with enriched diagnostic context.

This example illustrates a key property: GSI-LR's conservatism means it sometimes *cannot* improve in a given round, but it never makes things worse. The archived diagnostics from failed branches inform subsequent proposals, enabling eventual resolution when the right combination of edits becomes available.

## F FULL DEVELOPMENTAL TRACE

Table 5 presents a condensed developmental trace from seed 42, and Table 6 extends it with additional rounds.

Round	Event	Acc	FCR $\downarrow^\dagger$	Outcome
0	Initial state: 6/20 rules wrong	0.670	0.700	Baseline
3	Propose: flip $r_{14}$	—	—	<b>Rejected</b> (Step 3): fixes $F_7$ , breaks $F_{22}$ , $\Delta\text{FCR} = 0$ . RB +1.
7	Propose: flip $r_5 + r_{14}$ (compound, explorer branch)	0.680	0.675	<b>Accepted</b> : compound edit, $\Delta\text{FCR} = -0.025$ . Commit.
12	Propose: flip $r_8$	—	—	<b>Rejected</b> (Step 5): unauthorized evaluator change. RB +1.
50	Final state	0.691	0.675	DRR = 0.0, RB = 25.9

$^\dagger$ Lower FCR = fewer family contradictions = better.

Table 5: Condensed developmental trace (seed 42). The selected rounds illustrate the three canonical outcomes: rejection at Step 3, acceptance of a compound edit, and rejection at Step 5.

## G WORKED DCE METRIC TABLE

To demonstrate that DCE metrics are concretely computable, we present a worked example for Family  $F_7$  (5 members sharing rules  $\{r_2, r_5, r_{14}\}$ ).

Round	Event	Acc	FCR $\downarrow^\dagger$	Outcome
0	Initial state: 6/20 rules wrong	0.670	0.700	Baseline
1	Propose: flip $r_{12}$	—	—	Rejected (Step 3): $\Delta\text{FCR} = 0$
2	Propose: flip $r_{17}$	—	—	Rejected (Step 3): breaks $F_{15}$
3	Propose: flip $r_{14}$	—	—	Rejected (Step 3): fixes $F_7$ , breaks $F_{22}$
4–6	Various single-rule flips	—	—	All rejected (Step 3)
7	Compound: flip $r_5 + r_{14}$	0.680	0.675	<b>Accepted:</b> $\Delta\text{FCR} = -0.025$
8–11	Various proposals	—	—	All rejected (Step 3 or Step 4)
12	Propose: modify evaluator	—	—	Rejected (Step 5): unauthorized
13–49	Various proposals	—	—	All rejected
50	Final state	0.691	0.675	DRR = 0.0, RB = 25.9

$^\dagger$ Lower FCR = fewer family contradictions = better.

Table 6: Extended developmental trace (seed 42). The system accepts very few edits (AP = 0.007) because most proposals fail the FCR reduction criterion. The compound edit at round 7 is the critical accepted edit that moves FCR from 0.700 to 0.675.

### Family $F_7$ members and expected relations.

- $q_7^{\text{orig}}$ : Uses rules  $r_2, r_5, r_{14}$  with coefficients (1, 0, 1). Ground truth: True.
- $q_7^{\text{para}}$ : Same rules, coefficients (0, 1, 1) (2 flipped). Expected relation: *same* answer. Ground truth: True.
- $q_7^{\text{neg}}$ : Same rules, coefficients (1, 1, 1) (1 flipped). Expected relation: *opposite*. Ground truth: False.
- $q_7^{\text{impl}}$ : Rules  $r_2, r_5, r_{14}, r_3, r_{18}$  with extended coefficients. Expected relation: *same* as original. Ground truth: True.
- $q_7^{\text{perm}}$ : Same as original (reordered). Expected relation: *same*. Ground truth: True.

**Answers at round 0 (initial state, seed 42).** The policy has  $r_{14}$  wrong (believes True, ground truth False) and  $r_5$  wrong.

	$q_7^{\text{orig}}$	$q_7^{\text{para}}$	$q_7^{\text{neg}}$	$q_7^{\text{impl}}$	$q_7^{\text{perm}}$
<b>Round 0 answer</b>	True	True	True	True	True
<b>Ground truth</b>	True	True	False	True	True
<b>Correct?</b>	✓	✓	×	✓	×
<b>Round 50 answer</b>	True	True	False	True	True
<b>Correct?</b>	✓	✓	✓	✓	✓

Table 7: Family  $F_7$  answers at round 0 and round 50.

### DCE metrics for Family $F_7$ .

- **FCR (round 0):** The negation variant  $q_7^{\text{neg}}$  should have the opposite answer from  $q_7^{\text{orig}}$ , but both answer True. Family  $F_7$  is contradicted. Contribution to global FCR:  $1/40 = 0.025$ .
- **FCR (round 50):** After the compound edit at round 7 (flip  $r_5 + r_{14}$ ),  $q_7^{\text{neg}}$  now correctly answers False. The expected relation (opposite of  $q_7^{\text{orig}}$ ) is satisfied. Family  $F_7$  is no longer contradicted. Contribution to global FCR:  $0/40 = 0$ .
- **AP for  $F_7$ :** The edit at round 7 resolved  $F_7$ 's contradiction and never regressed through round 50. AP = 1.0 for this family's accepted edit.
- **DRR for  $F_7$ :** No regression occurred after the round-7 edit. DRR = 0.0.

This worked example demonstrates that DCE metrics are mechanically computable from the developmental trace and the family structure, requiring only the policy's answers at each round and the expected inter-question relations.

## H ILLUSTRATIVE CASE STUDIES

We present three scenarios illustrating the kinds of developmental failures GSI-LR is designed to prevent, grounded in the experimental traces.

**Example 1** (Quantifier-scope inconsistency). *Consider a family containing “Every reviewer read some paper” (distributed reading) and “There exists a single paper that every reviewer read” (shared reading). A local repair that improves handling of the first sentence by adding a scope-sensitive prompt may inadvertently cause the system to treat both sentences as equivalent — an edit that looks correct on one item but creates a family contradiction. In GSI-LR, the TCG would flag both sentences as members of a negation/non-equivalence family, and the validation cascade would reject any edit that makes them agree when they should not.*

**Example 2** (Cascading belief update). *This scenario mirrors the round-3 rejection in Table 5. The system incorrectly handles rule  $r_{14}$  (believes True, ground truth False). A critic identifies this error and proposes a fix. The fix correctly resolves Family  $F_7$  but breaks Family  $F_{22}$ , which also depends on  $r_{14}$  through an implication chain. The net  $\Delta\text{FCR} = 0$ : one family fixed, one family broken. Without family-aware governance, the fix passes (accuracy improves). With GSI-LR, Step 3 of the cascade computes  $\Delta\text{FCR}$  across all affected families and rejects the edit because  $0 > -\tau_c$ .*

**Example 3** (Branch-diverse diagnosis). *This scenario mirrors the round-7 acceptance in Table 5. Family  $F_7$  shows a persistent contradiction. A champion branch proposes flipping  $r_{14}$  alone (fixes  $F_7$ , breaks  $F_{22}$ ;  $\Delta\text{FCR} = 0$ , rejected). An explorer branch proposes flipping both  $r_5$  and  $r_{14}$  simultaneously — a compound edit that reduces  $F_7$ ’s contradiction without breaking  $F_{22}$  because the second flip compensates. The compound edit achieves  $\Delta\text{FCR} = -0.025$  and is accepted. Without branch diversity, the system would repeatedly try single-rule flips and fail.*

## I DEDUCTION, ABDUCTION, AND INDUCTION UNDER ONE ARCHITECTURE

The ICLR workshop call treats deduction, abduction, and induction as distinct reasoning modes (Dalvi et al., 2021; Creswell et al., 2022). In GSI-LR, the modes differ primarily in the kinds of commitments they produce and the validators they must satisfy, but they share the same governance architecture.

**Deduction.** A deductive branch starts from explicit premises and aims to derive or refute a target claim. Admission emphasizes proof validity, solver agreement, and invariance under paraphrase and premise permutation. A deductive edit that improves accuracy by weakening proof faithfulness should not be admitted.

**Abduction.** An abductive branch begins with an observed contradiction and proposes explanatory hypotheses. Abductive admission requires minimality of assumptions, compatibility with existing commitments, and non-contradiction under nearby family variants. The TCG is especially useful here because it stores not only accepted explanations but rival hypotheses and their disconfirming evidence.

**Induction.** Inductive branches search across repeated failure patterns and propose reusable repair rules. Induction is attractive because it converts many local failures into one policy change, but dangerous because it invites overgeneralization. Accordingly, inductive admission emphasizes cross-family robustness and maintenance cost.

The same contradiction cluster should often be attacked by multiple modes in parallel. A deductive branch might attempt a stronger proof decomposition; an abductive branch might search for a hidden premise; an inductive branch might propose a new translator rule. Over time, the system acquires a portfolio of lineages — some proof-specialized, some explanation-specialized, and some repair-rule-specialized.

## J FAMILY TYPES AND GENERATION TEMPLATES

**Trajectory reporting.** Every DCE evaluation should report: (1) an *edit ledger* documenting proposals, rejections, and rollbacks per tier; (2) a *contradiction-topology report* identifying which

Family type	Expected relation	Generation template
Negation	Predictable reversal ( $q$ and $\neg q$ )	Flip one premise coefficient; expected answer inverts
Paraphrase	Equivalent truth value	Flip 2 coefficients symmetrically; majority answer preserved under correct interpretation
Implication chain	Transitive consistency ( $A \rightarrow B \rightarrow C$ , test all links)	Add 2 superset premises; entailment preserved
Premise permutation	Same entailment outcome	Reorder premises identically; expected answer invariant
Contrapositive	$A \rightarrow B$ iff $\neg B \rightarrow \neg A$	Negate conclusion and swap direction; expected answer preserved

Table 8: Family types for DCE with explicit generation templates. Each type is defined by an expected logical relation and a reproducible construction rule.

family types remain hardest; (3) *FCR trajectories* over edit rounds with confidence bands; and (4) the improvement budget (solver calls, branch count, memory retention policy).

## K FALSIFIABLE HYPOTHESES

**H1.** Compared with local self-refinement, GSI-LR should reduce FCR at matched or better accuracy. If not, temporal contradiction memory and governance are unnecessary overhead.

**H2.** Compared with unconstrained search, GSI-LR should improve  $AP@k$  and lower DRR. If not, the validator does not distinguish durable edits from unstable ones.

**H3.** Solver-admissible editing should matter most on tasks requiring formal translation. If gains are uniformly weak, autoformalization is the bottleneck.

**H4.** Branch diversity should outperform single-trajectory search under the same edit budget when failures admit multiple plausible diagnoses.

## L CLAIM-TO-EVIDENCE MAPPING

Claim	Evidence	Strength
Uncontrolled search increases contradictions despite higher accuracy	UnconstrSearch vs. Static: FCR 0.775 vs. 0.740	Strong
DCE captures trajectory-level failures missed by snapshot accuracy	Developmental traces, 7 metrics	Strong
Solver-backed validation controls debt and non-regression	\Solver ablation: MD 4.931 vs. 0.467	Strong
Full governance stack (TCG, branches, edit-rights) improves consistency	Ablations show minimal FCR impact at this scale	Moderate*

\*Expected to strengthen in richer domains with delayed contradictions and evaluator self-modification.

Table 9: Claim-to-evidence mapping. The framework’s primary empirical support comes from the evaluation lens and solver validation; the full governance architecture is validated conceptually and expected to show stronger empirical signal at scale.

## M WHY GSI-LR OVER STATIC

Static achieves an FCR of 0.740 “for free” by doing nothing — it is a floor, not a strategy. Static cannot improve: it permanently locks in all 6 initial rule errors and every family contradiction they

1080 cause. GSI-LR actively searches for improvements while maintaining a consistency guarantee that  
1081 Static gets passively. In our experiments, GSI-LR reduces FCR from 0.700 (initial) to 0.675 —  
1082 finding edits that fix errors without cascading contradictions. The DRR = 0.0 for GSI-LR means zero  
1083 delayed regressions: every accepted edit was durably beneficial. By contrast, LocalRefine shows  
1084 what happens when one tries to improve *without* governance: accuracy drops from 0.689 to 0.599,  
1085 and DRR = 0.133 means 13.3% of accepted edits later cause regressions. In domains where the initial  
1086 policy is weak, Static locks in all initial errors permanently; GSI-LR can fix those errors that do not  
1087 cascade, and its governance prevents the cascading failures that plague LocalRefine.

1088 In realistic (non-deterministic) settings with stochastic LLM reasoning, unconstrained search would  
1089 face additional cascading contradictions beyond what our clean domain reveals, making the governed  
1090 tradeoff even more favorable.

1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133