Lean Refactor: Multi-Objective Controllable Proof Optimization via Agentic Strategy Search

Jialin Lu[1], **Soonho Kong**[2], **Rodrigo Stehling**[1], **Kaiyu Yang**[3], **Zhangyang Wang**[4], **Weiran Sun**[1], **Wuyang Chen**[1]
[1]Simon Fraser University [2]Amazon Web Services [3]Apodex
[4]University of Texas at Austin

Abstract

We present **Lean Refactor**, a plug-and-play retrieval-augmented agentic framework for *multi-objective, controllable, and version-robust* refactoring of Lean proofs. LLM-generated proofs are notoriously correct-but-verbose and brittle across library versions, yet existing refactoring works overlook **three practical challenges**: **1)** Lean refactoring is natively multi-objective (proof length, compilation cost, and version compatibility are often in tension); **2)** Lean repositories have fragile compatibility, whereas LLM releases are unaware of Lean/Mathlib versions; **3)** Training-based pipelines require repeated fine-tuning with each new LLM release, scaling neither with model churn nor with Lean's release cycle. Lean Refactor steers a **frozen** agentic LLM with retrievals from a curated database of **multi-objective refactoring strategies**, each densely annotated with metadata such as supported Lean/Mathlib versions and expected compilation-cost reduction. Experiments show over **70% token-level compression** on competition benchmarks, over **20%** on research repositories, and **up to 60% compilation-time reduction**, outperforming publicly released results from prior work and Claude Code. Version-filtered retrieval further improves compression on the target Lean version, and refactored miniF2F proofs exhibit stronger **zero-shot version transfer** to future Lean releases than their unrefactored counterparts. Project website: `https://leanrefactor.github.io/`.

1 Introduction

Lean and Mathlib [13, 29] enable collaborative, machine-verified mathematics and provide a strong reinforcement learning testbed that grounds large language model (LLM) reasoning in formal proofs, supporting gold-medal International Mathematical Olympiad (IMO) performance [43, 2] and rapid progress in LLM-powered provers [4, 46, 47, 37, 23, 24]. However, LLMs trained with reinforcement learning (RL) often produce Lean proofs that are correct but **notoriously long and hard to read** [3, 18]. Because RL rewards target proof correctness, models are insensitive to redundant steps or heavy automation where simple steps suffice. Such proofs are difficult for humans to understand and computationally expensive to compile, posing particular problems for large libraries like Mathlib.

Recent work on LLM-based code refactoring [15, 17] pursues either fine-tuning LLMs [18, 48, 16] or agentic frameworks built on pretrained general-reasoning LLMs [3, 14, 31, 35, 32, 20, 19, 12, 7, 52, 45]. However, LLMs have seen limited adoption for refactoring Lean proofs, and we point out **three key limitations**. ***First, Lean refactoring is natively multi-objective.*** Lean code is often simultaneously optimized along several axes, such as proof length and compilation cost. These objectives are entangled and frequently in tension: shorter proofs may invoke heavier tactics that inflate compilation cost, while cheaper proofs may sprawl in length. Standard pretraining, RLHF, and fine-tuning [33, 1] offer no principled mechanism to balance such trade-offs, let alone the **unseen objectives** users routinely introduce at inference time.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Second, frontier LLMs are misaligned with the Lean/Mathlib release cycle. Lean and Mathlib evolve on a near-weekly basis, with lemmas renamed, APIs restructured, and new automation landed, while LLMs are pinned to a knowledge cutoff. Even with web-search augmentation, **no principled way exists for maintainers to pair a codebase version with a compatible model**, leading to pervasive hallucination of stale tactics and removed APIs.

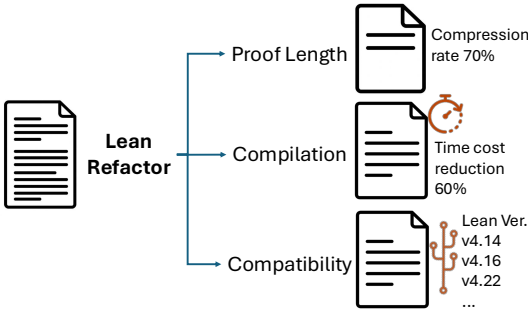


Figure 1: Our Lean Refactor can be controlled to refactor Lean proofs for multiple objectives: proof length, compilation time cost, and compatibility across Lean versions.

Third, paired refactoring data is scarce. Compared with mainstream languages (C/C++, Java, Python), Lean has a far smaller codebase, and existing Mathlib and community-aggregated competition corpora [53, 44, 22] provide neither refactoring trajectories nor noisy-clean code pairs.

To address these fundamental bottlenecks, we investigate our core question:

> *Can a frozen LLM be steered at inference time to balance competing refactoring objectives while reliably refactoring code in fast-evolving Lean/Mathlib environments?*

In this paper, we introduce **Lean Refactor**, a retrieval-augmented agentic framework that externalizes refactoring knowledge into a structured database. Lean Refactor grounds an agentic LLM with a curated corpus of **multi-objective, maintainable refactoring strategies**. Each retrieved strategy supplies the LLM with aligned structural context and explicit execution instructions (*when to apply*, *how to apply*, and *examples*). Crucially, these strategies are **densely annotated with refactoring metadata**, tracking signals such as expected compilation-cost reduction and validated Lean/Mathlib version compatibility. By dynamically filtering and reranking these annotated strategies at inference time, our framework steers proof optimization toward user-specified objectives without model fine-tuning, and remains version-robust as Lean/Mathlib evolves beyond the LLM's knowledge cutoff.

We summarize our contributions below:

1. We release the largest Lean refactoring dataset to date: over 200K long-short proof pairs and 9K refactoring strategies, each annotated with compilation-cost reduction and compatible Lean/Mathlib versions.
2. Lean Refactor achieves over 70% compression on competition proofs (miniF2F [53], PutnamBench [44], Putnam2025) and over 20% on research repositories (PFR [40], PhysLean [42], Analysis [39], FLT [9]), outperforming prior work and Claude Code while remaining compatible with Gemini, Claude, and GPT series backbones.
3. Lean Refactor supports multi-objective, controllable refactoring: beyond length, it cuts compilation cost by over 30%, holds compression steady across Lean versions, and produces miniF2F proofs that type-check across more future Lean/Mathlib releases than their unrefactored counterparts.

2 Lean Refactoring is Challenging for LLMs

Beyond the redundancy in code length demonstrated in previous papers [3, 18], in this section, we motivate our work with more practical challenges of Lean code refactoring with LLMs.

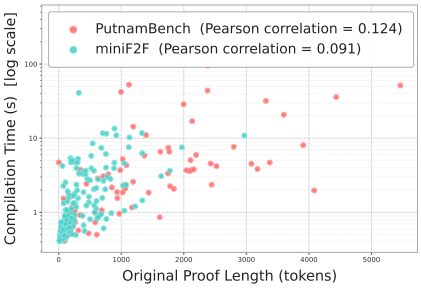


Figure 2: Weak correlation between proof length and compilation time. Token length fails to predict compilation cost because of short, heavy tactics, a factor often overlooked in previous Lean refactoring research.

Proof Optimization is Multi-Objective. While proof length is the most convenient refactoring objective, as it can be computed during LLM inference, compilation cost is an equally important complexity metric, and the two are not always aligned (Figure 2; see Appendix E for an example). Prior Lean refactoring work either ignores compilation cost or swaps it in as a substitute inference-time objective, rather than jointly optimizing it with length [3, 18]. More broadly, while recent formal-math post-training increasingly relies on verifiable rewards (e.g., compiler-judged correctness), continuous runtime signals such as **compilation cost remain underused** [24, 18].

Lean Repository Compatibility is Fragile. Version sensitivity is a long-standing bottleneck of Lean/Mathlib: Lean and Mathlib evolve rapidly; weekly releases frequently rename declarations and break imports even across patch versions [6], forcing downstream projects to pin to Mathlib's exact toolchain [27][1]. Agentic LLMs, however, are **never aligned** with a specific version: as Figure 3 shows, every recent frontier model's training cutoff lands *before* recent Mathlib releases, leaving newer APIs unseen and the model's coverage of older versions opaque. This pushes maintainers into continuous, costly upgrades. **Lean Refactor** mitigates this misalignment by grounding the agent in a strategy bank whose strategies carry explicit Mathlib-version metadata: retrieval is filtered to the user's target toolchain, and the bank itself can be re-profiled as new Mathlib versions ship.

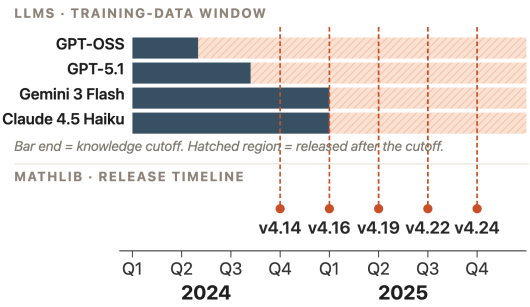


Figure 3: Knowledge cutoff lag in LLMs: Most LLMs misalign with Lean/Mathlib versions due to static knowledge cutoff.

Competitions Fail to Reflect Research-Level Proofs. Current LLM provers are primarily trained on competition mathematics, such as NuminaMath [22] and Goedel-Pset [23]. However, competition proofs do not reflect the structural complexity of research-level Lean code. As Table 11 shows, research proofs rely far more heavily on Mathlib and intra-project dependencies. This dense network of external and internal lemmas makes research-level code harder to refactor, forcing the model to reason over a much larger and less familiar context to find valid simplifications. Consequently, despite their practical importance, research-level proofs across diverse domains have been largely overlooked in prior Lean proof optimization work [18, 3].

3 Methods

Lean Refactor is a retrieval-augmented agentic framework built around three design choices that directly address the bottlenecks in Section 2. We externalize refactoring knowledge into a densely annotated strategy bank whose metadata, including Lean/Mathlib version compatibility and expected compilation cost reduction, mitigates the misalignment between the LLM's training cutoff and Lean/Mathlib release cycles (Section 3.1). At inference, the target proof is semantically grounded via dependency annotations [2], while a multi-objective retrieve-and-rerank pipeline steers the agent toward user-specified trade-offs between proof length, compilation cost, and version compatibility (Section 3.2). We wrap these components in a plug-and-play agent loop whose planner, refactorer, and debugger are instantiated by any frozen frontier LLM, so the framework transfers across Gemini, Claude, and GPT series without any retraining (Section 3.3). Figure 4 illustrates the pipeline; pseudocode is in Algorithm 1.

3.1 A Densely Annotated, Version-Aware Strategy Bank

Strategy structure. Each entry is a reusable refactoring pattern distilled from a raw long–short proof pair, organized into six fields: *title*, a natural-language *description*, a precondition pattern (*when to apply*), a step-by-step *application guide*, a *before/after example*, and a categorical *potential length reduction* tag. The schema preserves enough structural context for the LLM to generalize, promoting transfer to unseen proofs. Examples appear in Appendix Q.

Refactoring metadata. Each strategy is additionally annotated along two axes that the LLM cannot infer from its training data:

- *Compilation-time reduction.* We profile both proofs in every long–short pair with `lake env lean -profile`, isolating proof execution from import overhead, and record the relative speedup. Since each strategy can be instantiated by many long–short pairs, we keep the median speedup across all pairs associated with the strategy as the strategy-level estimate.

[1] In a Zulip thread on mathlib-dependent projects, the requirement is stated bluntly: "It needs to be the exact same version." See this discussion.

[2] We use adapted versions of jixia and ntp-toolkit for dependencies extraction.

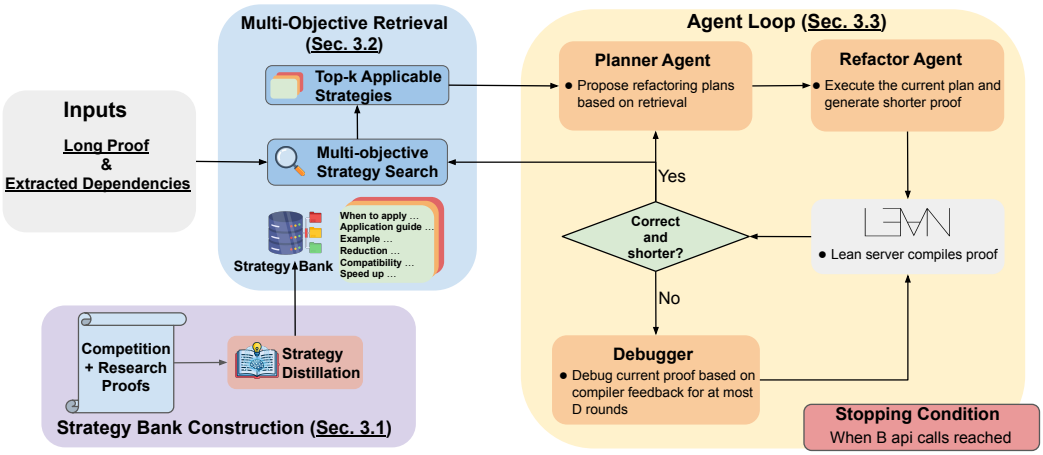


Figure 4: Overview of Lean Refactor framework. 1. We first summarize raw long-short proof pairs sourced from diverse Lean code sources into well-structured reusable strategies, and more importantly, we annotate each strategy with code refactoring metadata, including compilation time cost and version compatibility (Section 3.1). 2. We then train a strategy retrieval model that maps Lean code (that can be potentially optimized) to one or more refactoring strategies, with multi-objective re-rank and filtering (Section 3.2). 3. Finally, we design an agent workflow that intelligently plans refactoring, generates shortened proofs, and debugs (Section 3.3). Components in our agent workflow (planner, refactor, debugger) are frontier LLMs.

- *Version compatibility.* For all non-Mathlib pairs (Mathlib is locked to its toolchain), we recompile each shortened proof under three additional Lean toolchains beyond the original v4.24.0: v4.14.0, v4.16.0, and v4.22.0. A strategy's compatibility set is the intersection of toolchains under which all of its shortened proofs continue to compile, certifying that the transformation is robust across the interval rather than merely syntactically plausible in a single snapshot. This metadata is what makes the bank version-aware at inference time (Section 3.2) and is the primary lever by which we mitigate the misalignment between the LLM's training cutoff and Lean/Mathlib release cycles.

Bank construction. We aggregate statements from NuminaMath-1.5 [22], FineLeanCorpus [34], Mathlib, and ATLAS [25], and synthesize proofs with Goedel-Prover-V2-32B [24] and GPT-OSS-120B [30]. We decontaminate the aggregated theorems against our held-out evaluation set (defined in Section 4) via exact-match and embedding-based filtering to preclude leakage. Pairs are built by shortening long proofs and synthetically expanding short ones. GPT-OSS-120B distills each pair into one or more location-grounded strategies; a second LLM-as-judge pipeline filters incorrect extractions; and a Qwen3-Embedding-8B [51] based iterative de-duplication pipeline clusters semantically equivalent strategies. The final corpus contains **9237 unique strategies** distilled from 481567 raw strategy extractions over 200K long–short proof pairs. Full details are in Appendix C.

3.2 Multi-Objective Retrieval for Controllable Refactoring

Retrieval model training. We fine-tune Qwen3-Embedding-8B [51] on 639090 (query, strategy) pairs. Our training procedure largely follows prior Lean code search work [26], with several adaptations tailored to our refactoring setting. During strategy distillation from long–short proof pairs, we record the code segment in the long proof on which the strategy operates (Appendix C.4); we use this segment as the query and the strategy's *when-to-apply* clause as its positive target. Let b denote the mini-batch size, q_i the i-th query, c_i^+ its ground-truth positive strategy, and $\mathcal{C}_b = \{c_j^+\}_{j=1}^b$ the set of positives in the batch, where each c_j^+ with $j \neq i$ serves as an in-batch negative for query q_i. We write $\mathrm{Sim}(\cdot,\cdot)$ for cosine similarity, τ for the softmax temperature, and m for a false-negative margin. The indicator $\mathbb{I}(q_i, c)$ equals 0 when $\mathrm{Sim}(q_i, c) > \mathrm{Sim}(q_i, c_i^+) + m$, masking out candidates that score higher than the positive by more than m, and 1 otherwise. The training objective is

$$\mathcal{L} = -\frac{1}{b}\sum_{i=1}^{b} \log \frac{\exp(\mathrm{Sim}(q_i, c_i^+)/\tau)}{\sum_{c\in\mathcal{C}_b} \mathbb{I}(q_i, c)\, \exp(\mathrm{Sim}(q_i, c)/\tau)}. \tag{1}$$

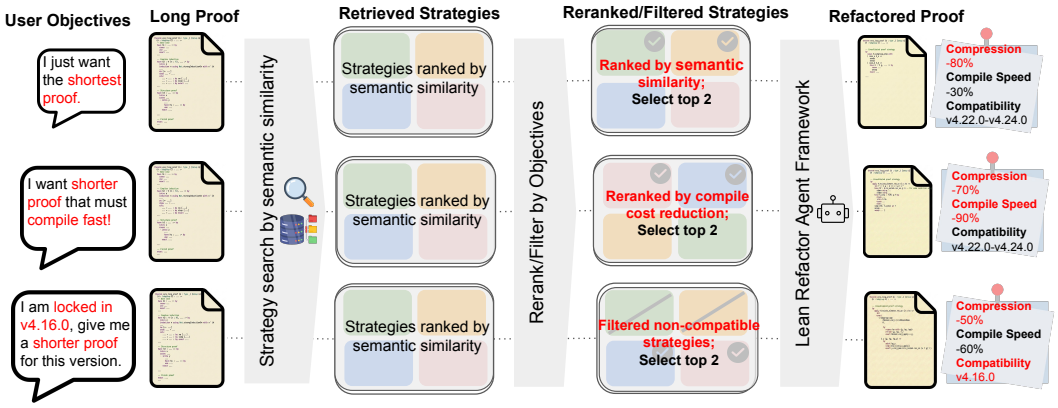


Figure 5: Multi-objective Lean proof refactoring. A single strategy bank adapts to diverse user objectives at inference without retraining. *Top:* for shortest proofs, top-K strategies by cosine similarity are used directly. *Middle:* to balance faster compilation, candidates are reranked by annotated compile-cost reduction. *Bottom:* for a target toolchain (e.g., v4.16.0), incompatible strategies are filtered out. The agent outputs (right) reflect these distinct trade-offs. Metadata-driven objectives are fully composable and easily extensible to new criteria.

We additionally apply boundary augmentation, jittering segment boundaries to bridge the gap between clean training spans and noisy runtime inputs. Training details are in Appendix G.

Objective-conditioned retrieval at inference (Figure 5). The user specifies an objective at inference time, and we reshape the retrieval rule accordingly while operating over the *same* underlying bank:

- *Shortest proof.* If the user prioritizes proof length, we use pure cosine-similarity top-K, exposing the agent to the most structurally relevant strategies.
- *Compilation-time aware.* If the user prioritizes compile-time reduction, we apply a two-stage retrieve-and-rerank procedure: we fetch a broad pool of semantically similar candidates, then rerank in descending order of annotated compile-cost reduction. Because the context window is bounded, top-K by similarity alone might exclude high-efficiency strategies; reranking surfaces them without sacrificing applicability.
- *Version-specific.* If the user targets a particular Lean version, we filter the candidate pool to strategies whose compatibility set includes that version, removing version-incompatible strategies.

Each objective is realized as a different retrieval rule over the same shared bank. This design composes cleanly: version filtering can be stacked on top of compile-time reranking to jointly target both objectives. It also admits *new* objectives at inference time by introducing new metadata fields without touching the model.

3.3 A Plug-and-Play, LLM-Agnostic Agent Loop

Iterative agent loop (Algorithm 1, Figure 4; see Appendix R for prompts). Given a long proof, we annotate it with the signatures of all invoked dependencies to provide semantic grounding. The annotated proof then enters an agentic loop that coordinates four distinct stages:

1. *Retrieval.* We segment the annotated proof at multiple granularities to enable refactorings of varying scope. Each segment is embedded and used to query the strategy bank; the top-K strategies are retrieved and processed under the user-specified objective (Section 3.2), then passed forward with their target locations.
2. *Planner.* A frozen LLM consumes the current proof, the retrieved strategies, and execution/planning history, and decomposes the task into a sequence of refactoring steps mapped to specific segments (or the entire proof).
3. *Refactorer.* The same LLM, given one planned step, rewrites the targeted segment (or the entire proof). The candidate is handed to the Lean 4 compiler for validation.

4. *Debugger.* On compile failure, compiler feedback is routed back to the LLM for up to D rounds of localized repair, constrained to the planner's original intent.

On successful compilation *and* length reduction, the updated proof and trace are fed back to the planner for dynamic replanning; on persistent failure, the step is logged and skipped. The loop terminates at a budget of B LLM calls or when proof length falls below target T. Appendix F describes configurable scalarized acceptance beyond this default length-only setting and reports the resulting length–compilation-cost trade-off.

Adaptability and Sustainability. Our design choices keep the framework agnostic to both LLM backbones and Lean versions. The planner, refactorer, and debugger are model-agnostic agent roles that operate over any frozen frontier LLM, so the framework can transfer to a new backbone with no fine-tuning or adaptation step. Refactoring knowledge, including strategies together with their version and compilation metadata, lives in an external bank that is retrieved at inference time and can be updated as Lean/Mathlib evolves. New optimization objectives are supported by acquiring the corresponding metadata, with no retraining required. As a result, new frontier LLMs inherit the bank's full coverage immediately, and new Lean versions are supported by re-profiling the bank. Section 4 empirically validates interchangeable deployment across Gemini, Claude, and GPT-OSS.

4 Experiments

Benchmarks. We evaluate Lean Refactor on seven benchmarks across two regimes. *Competition:* miniF2F [53] (194 theorems), PutnamBench [44] (75 theorems), and Putnam2025 (66 theorems from problems B1/B2 of Seed-Prover 1.5 [10]). For miniF2F and PutnamBench, we reuse the long proofs released by ProofOptimizer [18] so both methods refactor identical inputs. *Research:* 45 theorems sampled from each of Analysis [39], FLT [9], PFR [40], PhysLean [42], and PDE [38], along with our 15 generated solutions to Verina [50], spanning diverse proof lengths with all internal and external dependencies retained.

Setup and metrics. Unless noted otherwise, all runs use Gemini 3 Flash [41], Lean v4.24.0, an API call budget of 30, and the syntax-aware tokenizer from ProofOptimizer (code in Appendix M). We report three primary metrics: **proof length reduction** (Section 4.1), **compilation-time reduction** (Section 4.2), and **cross-version compatibility** (Section 4.3). Heartbeat reduction and additional compilation-cost results are deferred to Appendices N and O; dataset statistics, length distributions, and full experimental details to Appendix K.

Baselines. On miniF2F and PutnamBench, we compare against **ProofOptimizer** [18] using their released optimized proofs[3]; their closed-source checkpoints preclude evaluation on the remaining benchmarks. We additionally isolate each component of our framework via four internal ablations: **Base Agent** (refactorer + debugger + retrieval, planner removed), **Base Agent w/o Retrieval** (planner and retrieval removed), **Lean Refactor w/o Retrieval** (planner + refactorer + debugger, retrieval removed), and **Lean Refactor w/ Random Retrieval** (retrieval replaced by sampling over the bank).

To test LLM-agnosticism, we re-run our framework with Claude Haiku 4.5 [5] and compare against **Claude Code** (also powered by Haiku 4.5), giving Claude Code a dollar budget matched to our framework's Haiku 4.5 runs. We evaluate Claude Code under three tool configurations of increasing tool access: (i) Lean MCP[4] only, providing compiler feedback; (ii) Lean MCP + Lean skills[5], adding general Lean 4 refactoring and golfing guidance; and (iii) the above plus our strategy retrieval tool, which exposes the strategy bank as an on-demand resource the agent can query during refactoring. Due to budget, this comparison is restricted to one representative benchmark per regime: PutnamBench (competition) and Analysis (research-level). See Appendix K for further details.

[3] ProofOptimizer's released proofs are produced via an 8-iteration shortening process with sampling budgets of $6\times64 + 2\times1024$ per theorem (64 samples per iteration for the first 6, 1024 for the last 2) from a fine-tuned 7B model, while our system uses 30 API calls per theorem orchestrating planner, refactorer, and debugger steps over a frozen frontier LLM. The two paradigms differ in model scale and sampling-vs-agentic control flow, and since ProofOptimizer's checkpoints are not publicly available, a matched-backbone or matched-budget comparison is not possible; we therefore evaluate each system under its released configuration. Our ablations (Tables 1 and 2) isolate framework contributions on a shared backbone, and we additionally compare against Claude Code under matched budget and backbone (Table 3).

[4] `https://github.com/oOo0oOo/lean-lsp-mcp`

[5] `https://github.com/cameronfreer/lean4-skills`

4.1 Proof Length Reduction

Table 1: Average relative proof-length reduction (%) on competition benchmarks using Gemini 3 Flash. "–" indicates the baseline could not be evaluated. Best results in bold.

Method	miniF2F [53]	PutnamBench [44]	Putnam 2025
ProofOptimizer [18]	87.90	57.20	–
Base Agent	79.75	41.15	68.18
Base Agent w/o Retrieval	79.37	37.02	63.91
Lean Refactor w/o Retrieval	83.44	62.03	75.22
Lean Refactor w/ Random Retrieval	83.72	59.41	74.29
Lean Refactor	**88.40**	**70.38**	**79.27**

Table 2: Average relative proof-length reduction (%) on research-level benchmarks using Gemini 3 Flash. Best results in bold.

Method	Analysis [39]	FLT [9]	PFR [40]	PhysLean [42]
Base Agent	17.22	12.36	8.72	10.82
Base Agent w/o Retrieval	12.10	11.39	7.93	10.90
Lean Refactor w/o Retrieval	26.54	19.72	18.83	18.74
Lean Refactor w/ Random Retrieval	24.46	21.47	18.91	23.21
Lean Refactor	**33.54**	**23.55**	**20.37**	**27.69**

Tables 1 and 2 report length reductions with Gemini 3 Flash (test-time scaling curves in Appendix T); Table 3 reports the LLM-agnostic check with Haiku 4.5 and comparison with Claude Code. Additional results with PDE [38] and Verina [50] are shown in Appendix P. **Lean Refactor achieves state-of-the-art performance across both regimes:** on competition benchmarks it surpasses ProofOptimizer, with the gap widening on PutnamBench, where longer proofs benefit most from retrieval-augmented planning, and on human-written research-level proofs it achieves substantial additional compression and beats every ablation. **Planning and retrieval are synergistic.** The planner is critical because without it, the Base Agent becomes overwhelmed by long proofs and stalls in persistent debug loops. Conversely, retrieval only pays off when the planner explicitly sequences the provided strategies. Results show that random retrieval yields inconsistent improvements, whereas targeted retrieval delivers consistent and substantially larger gains. This confirms that the framework's success stems from *accurate strategy selection* rather than mere context augmentation. To quantify run-to-run variation, Appendix L reports three independent runs of targeted versus random retrieval. Because repeating every experiment was prohibitively costly, we focus on PutnamBench and FLT as representative competition and research benchmarks; in both cases, the gains are well outside the observed standard deviations.

Table 3: Average relative proof-length reduction (%) on PutnamBench and Analysis with Claude Haiku 4.5, compared with Claude Code. We evaluate our framework and ablation variants with Claude Haiku 4.5 as the backbone and compare against Claude Code (also powered by Claude Haiku 4.5) under three tool configurations: Lean MCP only, Lean MCP with Lean skills, and Lean MCP with Lean skills plus our strategy retrieval tool. Budget is matched by dollar cost. Best result in bold.

Agent	Configuration	PutnamBench [44]	Analysis [39]
Claude Code (Claude Haiku 4.5)	Lean MCP only	33.71	7.86
	+ Lean skills	33.29	8.83
	+ Our strategy retrieval tool	34.91	10.05
Ours (Claude Haiku 4.5)	Base Agent	39.21	15.45
	Base Agent w/o Retrieval	36.89	10.88
	Lean Refactor w/o Retrieval	48.76	18.33
	Lean Refactor w/ Random Retrieval	46.94	18.59
	Lean Refactor	**51.35**	**21.09**

The framework is LLM-agnostic and the strategy bank improves the performance of general-purpose agents. The same ranking holds when the backbone is switched to Claude Haiku 4.5 (Table 3), and under a matched budget, Claude Code itself improves on both benchmarks once given

access to our strategy retrieval tool. Even so, our full framework outperforms the strongest Claude Code configuration, demonstrating that our tailored planner and compiler-in-the-loop architecture provides targeted guidance that a general-purpose agent lacks, and more effectively leverages strategy retrieval. Beyond raw length, refactored proofs also reduce external and intra-project dependencies in 42–67% of cases across all benchmarks, evidence that the compression reflects genuine logical simplification rather than offloading work onto heavy library lemmas, which further improves readability and version resilience (Appendix H).

4.2 Multi-Objective Control: Compilation Time

Table 4: Compilation-time reduction on miniF2F, PutnamBench, and Putnam2025. *Avg. Rel. Reduc.*: average relative % decrease in compile time. *Avg. Abs. Reduc.*: average absolute decrease in seconds (higher is better for both). Each miniF2F/PutnamBench/Putnam2025 proof is compiled five times. Reported values are means across proofs; subscripts denote standard deviations across the five compilations. ProofOptimizer is not evaluated on Putnam2025 (model weights not publicly released). Best per dataset in bold.

Dataset	Method	Avg. Rel. Reduc. ↑ (%)	Avg. Abs. Reduc. ↑ (s)
miniF2F [53]	ProofOptimizer [18]	$-12.89_{\pm 0.55}$	$0.97_{\pm 0.05}$
	Lean Refactor w/o Retrieval	$26.56_{\pm 0.38}$	$3.02_{\pm 0.04}$
	Lean Refactor w/ Random Retrieval	$28.42_{\pm 0.48}$	$3.11_{\pm 0.05}$
	Lean Refactor	$-103.98_{\pm 2.64}$	$1.46_{\pm 0.06}$
	Lean Refactor w/ Reranking	$\mathbf{30.17}_{\pm 0.64}$	$\mathbf{3.14}_{\pm 0.04}$
PutnamBench [44]	ProofOptimizer [18]	$21.14_{\pm 0.43}$	$8.16_{\pm 0.07}$
	Lean Refactor w/o Retrieval	$36.32_{\pm 0.18}$	$7.05_{\pm 0.08}$
	Lean Refactor w/ Random Retrieval	$35.10_{\pm 0.35}$	$6.77_{\pm 0.07}$
	Lean Refactor	$37.11_{\pm 0.75}$	$8.97_{\pm 0.11}$
	Lean Refactor w/ Reranking	$\mathbf{42.62}_{\pm 0.27}$	$\mathbf{10.01}_{\pm 0.10}$
Putnam2025	Lean Refactor w/o Retrieval	$49.53_{\pm 0.32}$	$4.98_{\pm 0.08}$
	Lean Refactor w/ Random Retrieval	$50.13_{\pm 0.41}$	$4.99_{\pm 0.08}$
	Lean Refactor	$56.21_{\pm 0.34}$	$5.58_{\pm 0.06}$
	Lean Refactor w/ Reranking	$\mathbf{60.14}_{\pm 0.45}$	$\mathbf{6.05}_{\pm 0.06}$

We redirect the framework from length to compilation time by swapping only the retrieval rule (Section 3.2): **Lean Refactor w/ Reranking** reranks a semantically-similar candidate pool by each strategy's annotated compile-cost-reduction metadata. Controls are **w/o Retrieval**, **w/ Random Retrieval**, and the cosine-similarity-only **Lean Refactor**; ProofOptimizer is omitted on Putnam2025 (unreleased model weights). Compile time is measured via `lake env lean -profile`, excluding library import time. Table 4 reports average relative and absolute reductions on competition benchmarks; for research-level proofs, whose multi-file structure precludes clean per-proof wall-clock attribution, we instead measure heartbeats, Lean 4's deterministic proxy for elaboration work (Appendix O).

Lean Refactor w/ Reranking achieves the largest compilation-cost reduction on every competition and research-level benchmark (see Table 15 for research-level results), confirming that objective-aligned retrieval translates compile-cost metadata into wall-clock and heartbeat savings. miniF2F exposes the cost of objective *misalignment*: cosine-only retrieval yields a positive absolute reduction but a strongly negative relative reduction (-103.98%, i.e., compile time doubles on average in relative terms), driven by a few proofs whose compile time blows up by large factors, evidence that cosine-similarity-only retrieval can surface token-compressing yet compile-heavy strategies. Since ProofOptimizer's heartbeat-optimized proof variants are unreleased, we cannot measure their compilation time; therefore, Table 4 compares against their length-optimized proofs. However, under the heartbeat metric, **Lean Refactor w/ Reranking also beats ProofOptimizer's heartbeat-optimized numbers on miniF2F and PutnamBench** (Appendix N). More broadly, these results confirm the framework's adaptability: unlike models fine-tuned for a single objective, it is redirected across objectives by adjusting retrieval criteria over densely annotated metadata, with no retraining required.

4.3 Cross-Version Compatibility

The rapid release cycles of Lean and Mathlib make toolchain drift a recurring concern. We test (i) whether version-tagged retrieval metadata enables refactoring against arbitrary target Lean versions

and transfers across LLM backbones, and (ii) whether refactored proofs remain type-checkable as Mathlib evolves past their source version.

Version-aware retrieval across Lean releases. We restrict cross-version evaluation to PutnamBench: research-level projects are pinned to specific Lean/Mathlib toolchains through deep intra-project dependencies, so changing versions breaks compilation regardless of refactoring quality, whereas PutnamBench proofs are self-contained. The lack of version-portable research benchmarks is a notable limitation for evaluating compatibility. We evaluate three Lean versions (v4.14.0, v4.16.0, v4.22.0) under three retrieval settings: *No Retrieval*, *Full Retrieval* (all strategies), and *Filtered Retrieval* (strategies whose compatibility set covers the target version). Mathlib-derived strategies are excluded since Mathlib is locked to specific toolchains. We run both Gemini 3 Flash and GPT-OSS-20B; the latter is chosen because its knowledge cutoff predates *every* version we test (Figure 3), making it a clean test of whether version-filtered retrieval can supply toolchain-specific knowledge the backbone lacks. Results are reported in Table 5.

Table 5: Cross-version compatibility on PutnamBench. Average relative length reduction (%, higher is better). Best per (version, model) in bold. Budget is 30.

Version	Model	No Retrieval	Full Retrieval	Filtered Retrieval
v4.22.0	Gemini 3 Flash	55.99	55.43	**57.36**
	GPT-OSS-20B	27.56	27.80	**30.33**
v4.16.0	Gemini 3 Flash	60.61	60.41	**62.24**
	GPT-OSS-20B	27.11	28.37	**28.98**
v4.14.0	Gemini 3 Flash	58.51	**62.85**	59.54
	GPT-OSS-20B	**29.37**	29.14	27.58

Three findings emerge. **(1)** Filtered Retrieval attains the best length reduction on four of six (version, model) cells, and the pattern holds across both backbones. **(2)** Full Retrieval shows no consistent gain over No Retrieval, suggesting that unfiltered semantic neighbors can inject incompatible artifacts that disrupt refactoring. **(3)** On GPT-OSS-20B, Filtered Retrieval still beats both baselines at v4.22.0 and v4.16.0 despite the lack of pretraining exposure to those toolchains, providing direct evidence that the strategy bank compensates for missing in-weights knowledge. The exception is v4.14.0, where gains are less consistent. This is the only tested version inside Gemini 3 Flash's training window; Appendix A discusses how in-window version knowledge may make filtering less beneficial and describes version drift as an ecosystem-level challenge. To estimate run-to-run variation within our compute budget, we repeat the Gemini 3 Flash experiment three times at v4.22.0, the latest evaluated target version; Filtered Retrieval is separated from the baselines across these runs (Appendix L).

Zero-shot version transfer of refactored proofs. We further evaluate zero-shot version transfer of refactored proofs by recompiling v4.19.0 outputs from Lean Refactor, ProofOptimizer, and the original verbose proofs against Mathlib releases through v4.29.0, without further refactoring. On miniF2F, Lean Refactor leads from v4.22.0 onward and the gap widens at later releases (Table 6), showing that its refactored proofs continue to type-check across a longer span of future versions than both ProofOptimizer's proofs and the originals. On PutnamBench, Lean Refactor remains the strongest among proof optimization systems, though both shortened variants trail the originals, reflecting corpus-level differences in how proof length and version transfer interact. See Appendix I for full per-version results on both benchmarks.

Table 6: Zero-shot version transfer on miniF2F. Count of v4.19.0 proofs that still type-check at later releases without further refactoring. Best per row in bold. See Table 10 for the full results.

Version	Lean Refactor	ProofOptimizer	Original
v4.20.1	194	194	194
v4.22.0	**189**	188	187
v4.26.0	**168**	162	141
v4.29.0	**166**	161	140

5 Related Work

LLM-based provers [37, 24, 10] optimize correctness via RL but leave length, compile cost, and version compatibility unsupervised. Prior Lean refactoring [3, 18] targets a single axis: ImProver needs per-objective example corpora, ProofOptimizer fine-tunes for length with compile cost demonstrated only as an alternative inference-time objective, and neither handles Lean/Mathlib version drift. General-purpose LLM code refactoring [15, 17, 20, 7, 19, 16] assumes paired data, stable APIs, and test-suite validity, conditions Lean inverts. Lean Refactor externalizes refactoring knowledge into a shared strategy bank with compile-time and version metadata, turning new objectives into retrieval rules rather than training runs, with a model-agnostic, training-free agent loop. Extended related work is provided in Appendix B.

6 Limitations

Cross-version evaluation is limited to PutnamBench due to toolchain-pinned research projects, and strategy-level metadata uses conservative cluster-level aggregation rather than marginal attribution; see Appendix D for details.

7 Conclusion

We present Lean Refactor, an LLM-agnostic framework that decouples Lean proof refactoring logic from the underlying model via a densely annotated strategy bank, steering frozen agentic LLMs to balance proof conciseness, compilation efficiency, and toolchain compatibility. Comprehensive evaluations across competition suites and research repositories show that metadata-aware retrieval is critical, and that the distilled strategies transfer across LLM backbones and even improve the general-purpose agent Claude Code. Ultimately, combining this retrieval mechanism with our specialized planner and compiler-in-the-loop architecture yields advanced refactoring capabilities that general-purpose coding agents struggle to match, laying the groundwork for more maintainable and version-resilient formal mathematical proofs.

Acknowledgment

This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award NERSC DDR-ERCAP0034682 and ASCR-ERCAP0031463. ZW is supported by NSF Award 2523383 (DMS AIMING).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Tudor Achim, Alex Best, Alberto Bietti, Kevin Der, Mathïs Fédérico, Sergei Gukov, Daniel Halpern-Leistner, Kirsten Henningsgard, Yury Kudryashov, Alexander Meiburg, et al. Aristotle: Imo-level automated theorem proving. *arXiv preprint arXiv:2510.01346*, 2025.

[3] Riyaz Ahuja, Jeremy Avigad, Prasad Tetali, and Sean Welleck. Improver: Agent-based automated proof optimization. *arXiv preprint arXiv:2410.04753*, 2024.

[4] AlphaProof and AlphaGeometry teams. AI achieves silver-medal standard solving international mathematical olympiad problems. `https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/`, 2024.

[5] Anthropic. Claude 4.5 haiku model card, October 2025. Accessed: 2026-04-22.

[6] Anne Baanen, Matthew Robert Ballard, Johan Commelin, Bryan Gin-ge Chen, Michael Rothgang, and Damiano Testa. Growing mathlib: maintenance of a large scale mathematical library. In *International Conference on Intelligent Computer Mathematics*, pages 51–70. Springer, 2025.

[7] Jiameng Bai, Ruoyi Xu, Sai Wu, Dingyu Yang, Junbo Zhao, and Gang Chen. Polo: An llm-powered project-level code performance optimization framework. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI). IJCAI*, pages 7319–7328, 2025.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[9] Kevin Buzzard and Richard Taylor. FLT. `https://github.com/ImperialCollegeLondon/FLT`, 2025.

[10] Jiangjie Chen, Wenxiang Chen, Jiacheng Du, Jinyi Hu, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Wenlei Shi, Zhihong Wang, Mingxuan Wang, Chenrui Wei, Shufa Wei, Huajian Xin, Fan Yang, Weihao Gao, Zheng Yuan, Tianyang Zhan, Zeyu Zheng, Tianxi Zhou, and Thomas Hanwen Zhu. Seed-prover 1.5: Mastering undergraduate-level theorem proving via learning from experience, 2025.

[11] Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, et al. Seed-prover: Deep and broad reasoning for automated theorem proving. *arXiv preprint arXiv:2507.23726*, 2025.

[12] Bowen Cui, Tejas Ramesh, Oscar Hernandez, and Keren Zhou. Do large language models understand performance optimization? *arXiv preprint arXiv:2503.13772*, 2025.

[13] Leonardo De Moura, Soonho Kong, Jeremy Avigad, Floris Van Doorn, and Jakob von Raumer. The lean theorem prover (system description). In *Automated Deduction-CADE-25: 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings 25*, pages 378–388. Springer, 2015.

[14] Kayla DePalma, Izabel Miminoshvili, Chiara Henselder, Kate Moss, and Eman Abdullah AlOmar. Exploring chatgpt's code refactoring capabilities: An empirical study. *Expert Systems with Applications*, 249:123602, 2024.

[15] Dhruv Gautam, Spandan Garg, Jinu Jang, Neel Sundaresan, and Roshanak Zilouchian Moghaddam. Refactorbench: Evaluating stateful reasoning in language agents through code. *arXiv preprint arXiv:2503.07832*, 2025.

[16] Jingzhi Gong, Rafail Giavrimis, Paul Brookes, Vardan Voskanyan, Fan Wu, Mari Ashiga, Matthew Truscott, Mike Basios, Leslie Kanthan, Jie Xu, et al. Tuning llm-based code optimization via meta-prompting: An industrial perspective. *arXiv preprint arXiv:2508.01443*, 2025.

[17] Jingzhi Gong, Vardan Voskanyan, Paul Brookes, Fan Wu, Wei Jie, Jie Xu, Rafail Giavrimis, Mike Basios, Leslie Kanthan, and Zheng Wang. Language models for code optimization: Survey, challenges and future directions. *arXiv preprint arXiv:2501.01277*, 2025.

[18] Alex Gu, Bartosz Piotrowski, Fabian Gloeckle, Kaiyu Yang, and Aram H Markosyan. Proofoptimizer: Training language models to simplify proofs without human demonstrations. *arXiv preprint arXiv:2510.15700*, 2025.

[19] Xinyi He, Qian Liu, Mingzhe Du, Lin Yan, Zhijie Fan, Yiming Huang, Zejian Yuan, and Zejun Ma. Swe-perf: Can language models optimize code performance on real-world repositories? *arXiv preprint arXiv:2507.12415*, 2025.

[20] Muhammed Abdulhamid Karabiyik. Refactorgpt: a chatgpt-based multi-agent framework for automated code refactoring. *PeerJ Computer Science*, 11:e3257, 2025.

[21] Lean Prover Community. Mathlib bump. Zulip Chat Archive, FLT stream, 2026. Accessed: 2026-09-24.

[22] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. `https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf`, 2024.

[23] Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, et al. Goedel-prover: A frontier model for open-source automated theorem proving. *arXiv preprint arXiv:2502.07640*, 2025.

[24] Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, Jiayun Wu, Jiri Gesi, Ximing Lu, David Acuna, Kaiyu Yang, Hongzhou Lin, Yejin Choi, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction, 2025.

[25] Xiaoyang Liu, Kangjie Bao, Jiashuo Zhang, Yunqi Liu, Yu Chen, Yuntian Liu, Yang Jiao, and Tao Luo. Atlas: Autoformalizing theorems through lifting, augmentation, and synthesis of data, 2025.

[26] Jialin Lu, Kye Emond, Kaiyu Yang, Swarat Chaudhuri, Weiran Sun, and Wuyang Chen. Lean finder: Semantic search for mathlib that understands user intents. *arXiv preprint arXiv:2510.15940*, 2025.

[27] mathlib4. Dealing with breakages from updating. `https://github.com/leanprover-community/mathlib4/wiki/Using-mathlib4-as-a-dependency#dealing-with-breakages-from-updating`, 2025. GitHub repository.

[28] Scott Morrison. The first official release of Lean 4. Lean Community Blog, September 2023. Accessed: 2026-09-24.

[29] Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. 2021.

[30] OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025.

[31] Azim Ospanov, Farzan Farnia, and Roozbeh Yousefzadeh. Apollo: Automated llm and lean collaboration for advanced formal reasoning. *arXiv preprint arXiv:2505.05758*, 2025.

[32] Khouloud Oueslati, Maxime Lamothe, and Foutse Khomh. Refagent: A multi-agent llm-based framework for automatic software refactoring. *arXiv preprint arXiv:2511.03153*, 2025.

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[34] Zhongyuan Peng, Yifan Yao, Kaijing Ma, Shuyue Guo, Yizhe Li, Yichi Zhang, Chenchen Zhang, Yifan Zhang, Zhouliang Yu, Luming Li, Minghao Liu, Yihang Xia, Jiawei Shen, Yuchen Wu, Yixin Cao, Zhaoxiang Zhang, Wenhao Huang, Jiaheng Liu, and Ge Zhang. Criticlean: Critic-guided reinforcement learning for mathematical formalization, 2025.

[35] Yonnel Chen Kuang Piao, Jean Carlors Paul, Leuson Da Silva, Arghavan Moradi Dakhel, Mohammad Hamdaqa, and Foutse Khomh. Refactoring with llms: Bridging human expertise and machine understanding. *arXiv preprint arXiv:2510.03914*, 2025.

[36] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020.

[37] ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

[38] Rodrigo Stehling, Jialin Lu, Wuyang Chen, and Weiran Sun. Lean formalization of pde topics. `https://github.com/weiran-sun/pde`, 1 2026. GitHub repository.

[39] Terence Tao. A Lean companion to Analysis I. `https://github.com/teorth/analysis`, 2024.

[40] Terence Tao. Formalization of the Polynomial Freiman-Ruzsa conjecture of Marton. `https://github.com/teorth/pfr`, 2025.

[41] The Gemini Team. Gemini 3 flash: frontier intelligence built for speed. `https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/`, December 2025. Accessed: 2026-03-16.

[42] Joseph Tooby-Smith. physlib: A project to digitalise results from physics into Lean. `https://github.com/leanprover-community/physlib`, 2025.

[43] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[44] George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition, 2024.

[45] Yue Wu, Minghao Han, Ruiyin Li, Peng Liang, Amjed Tahir, Zengyang Li, Qiong Feng, and Mojtaba Shahin. Fasterpy: An llm-based code execution efficiency optimization framework. *arXiv preprint arXiv:2512.22827*, 2025.

[46] Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

[47] Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, et al. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv preprint arXiv:2408.08152*, 2024.

[48] Jiuding Yang, Shengyao Lu, Hongxuan Liu, Shayan Shirahmad Gale Bagi, Zahra Fazel, Tomasz Czajkowski, and Di Niu. Perfcoder: Large language models for interpretable code performance optimization. *arXiv preprint arXiv:2512.14018*, 2025.

[49] Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36:21573–21612, 2023.

[50] Zhe Ye, Zhengxu Yan, Jingxuan He, Timothe Kasriel, Kaiyu Yang, and Dawn Song. Verina: Benchmarking verifiable code generation, 2026.

[51] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025.

[52] Yuwei Zhao, Yuan-An Xiao, Qianyu Xiao, Zhao Zhang, and Yingfei Xiong. Semopt: Llm-driven code optimization via rule-based analysis. *arXiv preprint arXiv:2510.16384*, 2025.

[53] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics, 2022.

A Interpreting the v4.14.0 Compatibility Result

Table 5 shows the least consistent benefit from compatibility filtering at `v4.14.0`. This is the only tested version that falls inside Gemini 3 Flash's training window (Figure 3), so the backbone may already have relevant version knowledge. Consistent with this, Filtered Retrieval still beats No Retrieval for Gemini even at `v4.14.0` (59.54 vs 58.51); the deviation is Full Retrieval overtaking both, which is consistent with the possibility that the backbone can repair version drift in retrieved strategies, allowing unfiltered retrieval to benefit from a larger pool of relevant strategies.

More broadly, version incompatibility is an acknowledged ecosystem-level bottleneck rather than an artifact of our setup. Lean's first stable-release announcement states that backward compatibility between versions cannot yet be promised [28]. Mathlib's dependency guide documents that declarations move or are renamed as the library evolves [27], and its maintainers describe breaking changes and the deprecation infrastructure required to manage them at scale [6]. The FLT project likewise maintains a long-running "Mathlib bump" discussion recording the recurring work needed to keep a major research formalization synchronized with Mathlib [21]. We are not aware of a community benchmark or infrastructure for evaluating LLM systems under version drift, as research projects typically pin exact toolchains [6].

B Extended Related Work

LLM-based theorem proving in Lean. A rapid line of work trains LLMs to generate Lean proofs end-to-end with reinforcement learning from compiler-verified rewards, including DeepSeek-Prover [46, 47, 37], Goedel-Prover [23, 24], Seed-Prover [11, 10], building on earlier tactic-level and retrieval-augmented systems [36, 49]. These rewards, however, primarily target proof correctness, leaving length, compilation cost, and version compatibility unsupervised. The resulting proofs are correct but verbose, slow to compile, and brittle across Mathlib releases [3, 18]. Our framework operates downstream of these provers, refactoring their outputs under multiple cost objectives without retraining the underlying model.

Lean proof optimization. Two prior systems directly target Lean 4 proof simplification. ImProver [3] wraps a black-box LLM in an agentic loop combining Chain-of-States prompting, best-of-n sampling with iterative refinement, and retrieval over Lean/Mathlib documentation and per-metric example databases. ProofOptimizer [18] instead fine-tunes a dedicated model on synthesized long-short pairs to directly shorten proofs. While both achieve length reductions, each restricts refactoring to a single-objective paradigm. ProofOptimizer's objective is frozen at training time. ImProver accommodates arbitrary user-defined metrics by using them to drive both example retrieval and candidate selection; however, scaling to new objectives requires constructing an entirely new, metric-specific example corpus. Crucially, neither system addresses the broader bottlenecks of Lean 4 proof maintenance identified in Section 2. Compilation cost is treated as a secondary objective swapped in at inference rather than jointly optimized [18], and continuous Lean 4 and Mathlib version drift remains invisible to both frozen fine-tuned models and standard example-based retrievers. Lean Refactor overcomes these limitations by retrieving *refactoring strategies* densely annotated with multi-objective execution metadata, such as expected compilation-time reductions and validated Lean 4 version compatibility. This enables Lean Refactor to turn complex, competing objectives into configurable retrieval rules over a single, shared strategy bank, bypassing the need for training-time commitments or fractured, per-metric datasets.

LLM-based code refactoring and performance optimization. A broader literature studies LLM-driven refactoring [15, 17] and code performance optimization for general-purpose languages. Agentic and prompting-based systems target maintainability and code quality [14, 35, 32, 20], while a parallel thread pursues runtime speedup [7, 19, 12, 52, 45, 48, 16]. These methods operate on settings such as Python, C++, and Java, where paired refactoring data is plentiful, APIs are comparatively stable, and correctness is verified by test suites rather than a strict type checker. Lean inverts each of these conditions: paired data is scarce, the toolchain breaks compilation weekly, and validity is a hard binary signal. Lean Refactor's agentic system is distinguished by the densely annotated strategy bank and a compiler-in-the-loop validation tailored to Lean's release dynamics.

C Details about Strategy Bank Construction

This section provides a complete description of the data pipeline summarized in Section 3.1. We detail (i) the source repositories from which formal statements are aggregated, (ii) the dual-strategy procedure used to synthesize aligned long–short proof pairs, (iii) the profiling and cross-version recompilation pipeline that produces the refactoring metadata, (iv) the prompting protocol used to distill location-grounded refactoring strategies from each pair, (v) the quality-control filter, (vi) the iterative embedding-based deduplication pipeline, and (vii) the procedure by which pair-level metadata is aggregated into strategy-level annotations.

C.1 Source Repositories of Lean Statements

To ensure that our strategy bank generalizes across both competition mathematics and a broad spectrum of undergraduate- and research-level domains, we aggregate formal statements from four complementary repositories.

- **NuminaMath-1.5 [22].** A large-scale corpus of formalized competition problems. NuminaMath provides a high density of challenging Olympiad-style Lean statements that exercise the full range of tactical machinery typically employed in contest proofs.
- **FineLeanCorpus [34].** A second large-scale source of competition-level statements, complementary to NuminaMath in style and topical coverage.
- **Mathlib4.** The standard library for the Lean 4 ecosystem. We extract theorems from `Mathlib4` to represent broader mathematical disciplines that are underrepresented in competition-oriented corpora.
- **ATLAS [25].** A synthetic dataset of theorem statements lifted from `Mathlib4` contexts. ATLAS broadens the stylistic distribution of statements beyond what is naturally present in human-curated corpora and serves as a useful augmentation for diversity.

This hybrid sourcing strategy exposes the downstream pipeline to a wide range of Lean coding styles, tactic preferences, and proof-architecture patterns. The held-out evaluation set used to assess generalization is described separately in the experiments section of the main paper.

Decontamination against the held-out evaluation set. To preclude leakage from our held-out evaluation set into the long–short proof pairs used to build the strategy bank, we apply a two-stage decontamination filter to every theorem statement before proof synthesis. First, we perform exact-match deduplication on normalized statements. Second, we perform embedding-based near-duplicate detection: each candidate training statement is encoded with Qwen3-Embedding-8B [51] and compared against the embeddings of all evaluation statements. If the top-1 cosine similarity is below 0.8, we treat the candidate as non-duplicate; if it is at or above 0.8, we invoke GPT-OSS-120B as an LLM-as-judge with the candidate and the matched evaluation statement to make the final determination, and discard the candidate if the judge confirms duplication. This filter is applied upstream of proof synthesis and strategy distillation, so the resulting 200K long–short pairs, the 481,567 raw strategy extractions, and the final corpus of 9,237 unique strategies are all derived exclusively from theorems that survive both stages.

C.2 Proof Synthesis and Long–Short Pair Construction

Proof generation. With the exception of `Mathlib4`, the sourced repositories provide only theorem statements without ground-truth proofs. We therefore synthesize formal proofs using two prover models: Goedel-Prover-V2-32B [24] and GPT-OSS-120B [30]. Using two provers rather than one is a deliberate design choice: it broadens the distribution of tactic patterns, naming conventions, and structural choices that subsequently appear in the long–short pairs and, by extension, in the distilled strategies.

For each theorem statement we sample between one and three candidate proofs per model, with the exact count bounded by our compute budget. To guarantee environmental consistency during verification, every candidate is prepended with a standardized header containing the necessary `import` directives and `namespace` declarations before being submitted to the Lean 4 compiler. When

multiple candidate proofs for the same theorem successfully verify, we retain the *longest* valid proof, since verbose proofs provide the richest foundation for refactoring and yield the most informative long–short pairs in the subsequent stage.

Aligned long–short pair construction. Given a corpus of verified proofs, we construct aligned pairs through a length-conditioned, dual-strategy prompting pipeline:

- **Forward simplification (long → short).** For proofs exceeding 50 tokens, we prompt GPT-OSS-120B to optimize the trajectory and produce a shorter counterpart.
- **Reverse complexification (short → long).** For proofs under 50 tokens, we prompt GPT-OSS-120B to synthetically expand the proof into a more verbose, less idiomatic counterpart, yielding the "long" element of the pair.

The reverse-complexification branch ensures that compact, idiomatic proofs, which would otherwise be excluded because no naturally occurring "long" counterpart exists, still contribute to the bank by surfacing the simplifications they implicitly embody.

To improve structural fidelity in both directions, we augment the prompts with explicit dependency information that enumerates the lemmas and premises invoked in the input proof. Empirically, this dependency context reduces hallucinated lemma references and produces transformations that are more faithful to the underlying proof structure. Each candidate transformed proof is recompiled to confirm that the produced pair consists of two independently verified Lean proofs of the same theorem.

In total this pipeline yields **200K verified long–short proof pairs**, which form the foundation for all subsequent strategy distillation.

C.3 Refactoring Metadata Annotation

Each long–short pair is annotated along two axes that the LLM cannot infer from its training data alone: compilation-time reduction and Lean-toolchain compatibility.

Compilation-time profiling. We measure the compilation time of both the long and the short proof in each pair using the `lake env lean -profile` command, which exposes fine-grained per-declaration timing. We explicitly isolate proof execution from library import overhead, since import time is dominated by caching effects unrelated to the proof content itself. Each measurement is taken in a consistent environment to ensure stable measurements, and we record the relative reduction in execution time achieved by the shortened proof. This per-pair speedup is the unit from which strategy-level estimates are later aggregated (Appendix C.8).

Cross-version compatibility testing. To assess whether a refactoring remains valid under future Lean releases, we recompile each shortened proof under multiple Lean and `Mathlib` toolchains beyond the original v4.24.0: `v4.14.0`, `v4.16.0`, and `v4.22.0`. We restrict version-compatibility analysis to non-`Mathlib` theorems, because `Mathlib` itself is tightly coupled to a specific Lean release via `lean-toolchain` and routinely fails to build under manual version overrides; this coupling makes any cross-version signal collected on `Mathlib` theorems an artifact of infrastructure rather than of the refactoring strategy under study.

For each eligible pair we track the compilation status of both the long and short proof under each toolchain. A refactored proof is deemed compatible with a target version if the short proof compiles successfully in that version. The strategy is considered robust on the interval spanned by the toolchains under which all of its shortened proofs continue to compile, a notion that is formalized at the strategy level in Appendix C.8.

C.4 Strategy Summarization with Location Grounding

A single long–short proof pair typically encompasses several distinct refactoring operations applied to different regions of the code. To capture this granularity, we prompt GPT-OSS-120B with both proofs and ask it to extract a set of applied refactoring strategies. Crucially, we instruct the model to ground each extracted strategy to a specific component of the original long proof, identified by start and end line numbers.

This grounding requirement serves two purposes:

1. *Hallucination control.* In preliminary experiments we observed that without an explicit anchoring constraint, the model frequently fabricated strategies that did not correspond to any concrete change between the long and short proofs. Forcing each strategy to point to a specific code region eliminates a large class of these hallucinations and improves extraction quality.
2. *Retrieval supervision.* The proof-component-to-strategy mapping is the supervision signal used to fine-tune the strategy-retrieval model employed by our agent at inference time. Given a proof component currently being refactored, the retriever returns the strategies historically associated with similar components, providing contextually relevant guidance.

Importantly, location grounding does not bias the extracted strategies toward purely localized fixes. Global structural refactorings, such as restructuring an entire case split or replacing a multi-step induction with a single library lemma, are accommodated by grounding the strategy to a correspondingly larger contiguous chunk of the proof.

C.5 Strategy Schema

To ensure clarity, applicability, and reusability, each distilled strategy is organized into the following fields. Together they form a self-contained record sufficient for the agent to reason about *when*, *why*, and *how* to apply the refactoring at inference time. Examples are provided in Appendix Q.

- **Title.** A short natural-language summary of the high-level idea of the refactoring strategy.
- **Description.** A natural-language statement of the core conceptual idea behind the refactoring. It includes information like the semantic shift that the strategy brings (e.g. replacing a manual step-by-step existence construction or a redundant case analysis with a more powerful library lemma or automated tactic).
- **When to Apply.** A generalized description of the Lean code pattern or logical situation that serves as a precondition for the strategy. This field explicitly identifies the structural "anti-patterns" that warrant the proposed refactoring and enables the agent to recognize applicable contexts in unseen proofs.
- **Application Guide.** A step-by-step, actionable refactoring guide. It provides the agent with deterministic, sequential instructions for executing the transformation, minimizing ambiguity at the point of application.
- **Abstract Example.** A generalized exemplar that supplies a concrete "before" and "after" snippet of the transformation. Specific variable names and domain artifacts are abstracted away, so that the example highlights the underlying syntactic and tactical shift rather than the surface details of any one theorem.
- **Potential Reduction.** A categorical estimate (*high*, *medium*, or *low*) of the expected impact on proof length.

C.6 Quality Filtering via LLM-as-Judge

To guarantee both the correctness of the distilled strategies and the fidelity of their location grounding, we apply a rigorous filtering stage. Each extracted strategy is evaluated by an LLM-as-judge alongside its corresponding long–short proof pair and the identified proof component. The judge assesses two criteria:

1. *Transformation correctness.* Whether the proposed strategy correctly and logically transforms the identified component of the long proof into its optimized counterpart in the short proof.
2. *Schema fidelity.* Whether all six fields of the strategy schema (Appendix C.5) are accurately and consistently populated, with no contradictions between, e.g., the *When to Apply* precondition and the *Abstract Example*.

Only strategies passing both checks are retained for the next stage. Strategies that fail either check are discarded; we do not attempt to repair them, as our preliminary experiments indicated that repair attempts frequently introduce subtler errors that propagate downstream.

C.7 Clustering and Iterative Deduplication

The filtered set still contains substantial redundancy: many distinct long–short pairs surface the same underlying refactoring idea, expressed with surface-level variation. To prevent redundant strategies from consuming valuable prompt context at inference time, and to keep retrieval efficient, we curate a compact, semantics-aware subset of unique strategies via the following iterative pipeline.

Initialization. We seed the unique-strategy set by prompting GPT-OSS-120B to identify a diverse initial collection of canonical strategies from a sampled subset of the pool. This produces a high-coverage starting point against which subsequent strategies can be compared.

Iterative deduplication. Each remaining strategy is then processed sequentially against the dynamically growing unique set via a four-stage cascade:

1. *Embedding and candidate retrieval.* We encode the strategy's *Description* field with the Qwen3-Embedding-8B [51] model and retrieve the top-10 most semantically similar entries from the current unique set by cosine similarity.
2. *High-confidence shortcut.* If the top-1 cosine similarity is ≥ 0.9, the strategy is treated as a duplicate of the matched entry without further inspection. We adopt this shortcut because manual inspection of borderline cases indicated that pairs above this threshold are essentially always paraphrases of the same underlying refactoring, and skipping the judge call at this end of the distribution yields a substantial reduction in pipeline cost without a measurable loss in cluster purity.
3. *Judge-based duplicate detection.* If the top-1 cosine similarity is < 0.9, we prompt GPT-OSS-120B as an LLM-as-judge with the current strategy together with the 10 retrieved candidates. The judge decides whether the current strategy is a semantic duplicate of any existing entry. This stage handles the long tail of paraphrastic variation that pure embedding similarity tends to underestimate (e.g., strategies that share an underlying tactical idea but use different terminology or describe it at different levels of abstraction).
4. *Update.* If either the shortcut or the judge identifies a duplicate, the current strategy is *clustered* under the matched existing entry, contributing its pair-level metadata to that cluster's aggregate (Appendix C.8). Otherwise, the strategy is appended to the unique set as a new canonical entry and becomes a candidate for matching against subsequent strategies.

By combining embedding-based shortlisting, a high-similarity shortcut, and an LLM judge fallback, this cascade keeps the procedure tractable on the full filtered pool while still benefiting from the judge's ability to resolve paraphrastic variation.

Final corpus. Applying this pipeline to the filtered pool of 481,567 raw extractions yields a refined corpus of **9,237 unique refactoring strategies**, each backed by a cluster of constituent long–short pairs that share its semantics.

C.8 Strategy-Level Metadata Aggregation

The pair-level metadata collected in Appendix C.3 is lifted to the strategy level by aggregating across the cluster of long–short pairs assigned to each unique strategy.

Compilation-time reduction. For each strategy, we keep the median relative execution-time reduction across all long–short pairs in its cluster to get a robust strategy-level estimate.

Version compatibility set. Following Appendix C.3, we restrict attention to non-`Mathlib` pairs. A strategy's *compatibility set* is defined as the intersection of toolchains under which the shortened proofs of all its constituent pairs compile successfully. This intersection-based definition is conservative by design: a strategy is certified for a toolchain only if every observed instantiation of the strategy survives recompilation under that toolchain. Aggregating across all constituent pairs, rather than relying on a single example, ensures that a toolchain enters the compatibility set only when the strategy has been shown to transfer reliably across diverse instantiations, rather than happening to compile in one isolated case.

Downstream use. The aggregated metadata is the lever that makes the bank *version-aware* and *efficiency-aware* at inference time: the multi-objective retrieval procedure (Section 3.2) uses the compatibility set to filter strategies incompatible with the user's target toolchain and the median speedup to rank surviving candidates. This is the mechanism by which we decouple refactoring reliability from the LLM's training cutoff.

D Limitations

Benchmark coverage for cross-version evaluation. Our cross-version compatibility experiments (Section 4.3) are confined to PutnamBench because, to our knowledge, no existing Lean benchmark supports compatibility evaluation beyond self-contained competition problems. Research-level projects pin to specific Lean/Mathlib toolchains through deep intra-project dependencies, so changing the toolchain breaks compilation regardless of refactoring quality. Constructing version-portable research-level benchmarks, e.g., proofs whose dependencies are at a granularity that decouples them from the project's pinned toolchain, is an important direction for the community.

Joint-effect attribution. Strategy-level compile-time annotations are computed as the cluster median of pair-level reductions, where each pair contributes the joint speedup of its refactorings; analogously, version-compatibility sets are intersected across constituent pairs, so a strategy inherits the toolchain coverage that all its instantiations jointly satisfy. Both choices fit the role the metadata plays at inference: the retrieval module reranks candidates by *relative* compile-time benefit and filters by guaranteed version coverage, for which the median and intersection are robust, conservative estimators. Disentangling marginal per-strategy contributions, via Shapley-style attribution, or regression over co-occurrence indicators, would further enrich the schema along a new axis and is a natural extension we leave to future work.

E An Example of Proof Length and Compilation Time Tradeoff

To illustrate the misalignment of proof length and compilation time, consider the two alternative Lean proofs for the theorem `amc12_2001_p21` from miniF2F [53] below.

```
theorem amc12_2001_p21
  (a b c d : ℕ)
  (h0 : a * b * c * d = Nat.factorial 8)
  (h1 : a * b + a + b = 524)
  (h2 : b * c + b + c = 146)
  (h3 : c * d + c + d = 104) :
  a - d = (10 : ℤ) := by
  norm_num [Nat.factorial] at h0
  have : b <= 525 := by nlinarith
  interval_cases b <;> simp_all
  <;> nlinarith
```

```
theorem amc12_2001_p21
  (a b c d : ℕ)
  (h0 : a * b * c * d = Nat.factorial 8)
  (h1 : a * b + a + b = 524)
  (h2 : b * c + b + c = 146)
  (h3 : c * d + c + d = 104) :
  a - d = (10 : ℤ) := by
  -- Factor hypotheses into (x+1)(y+1) =
    xy + x + y + 1
  have h4 : (a+1)*(b+1) = 525 := by
    -- proof omitted
  have h5 : (b+1)*(c+1) = 147 := by
    -- proof omitted
  have h6 : (c+1)*(d+1) = 105 := by
    -- proof omitted
  -- Pin down each variable via gcd +
    interval_cases
  have h7  : b = 20 := by -- proof omitted
  have h8  : a = 24 := by -- proof omitted
  have h9  : c = 6  := by -- proof omitted
  have h10 : d = 14 := by -- proof omitted
  -- Cast to integers and conclude
  have h11 : (a:ℤ) - (d:ℤ) = 10 := by
    -- proof omitted
  exact_mod_cast h11
```

The short proof is concise but relies on a single heavy cascade `interval_cases, simp_all, nlinarith` over the loose bound $b \leq 525$, so Lean's most expensive tactics fire on hundreds of

branches. Conversely, the second proof spans over 130 lines of explicit, step-by-step arithmetic deductions. Despite its much longer text, it significantly reduces the compiler's search burden and compiles more than $20\times$ faster than its shorter counterpart. This stark contrast highlights that optimizing LLMs solely for shorter text lengths can drastically penalize underlying runtime metrics.

F Multi-Objective Formulation and Candidate Acceptance

This section clarifies how Lean Refactor balances competing objectives and when it can retain a longer proof to reduce compilation cost. We formalize the multi-objective problem and its scalarized acceptance rule, report the length–heartbeat trade-off on PutnamBench under different objective weights, and describe constrained variants.

Objectives and scalarization. For a proof P of a fixed theorem, let $f_1(P)$ denote its token length, $f_2(P)$ its compilation cost, and $f_3(P)$ the number of incompatible toolchains in a fixed evaluation set of Lean/Mathlib versions. The refactoring problem has the vector objective

$$\min_{P \in \mathcal{P}_{\text{valid}}} \mathbf{F}(P) = \big(f_1(P), f_2(P), f_3(P)\big), \tag{2}$$

where $\mathcal{P}_{\text{valid}}$ contains proofs that pass verification in the target environment. We select a trade-off through weighted-sum scalarization, a standard approach to vector optimization [8, Section 4.7.4]:

$$\widehat{F}_{\mathbf{w}}(P) = \sum_{i=1}^{3} w_i f_i(P), \qquad w_i \geq 0, \quad \sum_{i=1}^{3} w_i > 0. \tag{3}$$

The weights specify how much improvement in one objective compensates for deterioration in another. Their values depend on the units used: for example, if compilation cost is measured in heartbeats, $(w_1, w_2, w_3) = (1, 0.01, 0)$ assigns the same cost to one token as to 100 heartbeats. Fixing the weights gives a single score for comparing candidate proofs under this chosen trade-off. It does not find all Pareto-optimal trade-offs. If compatibility with specified Lean/Mathlib toolchains is mandatory, candidates that fail verification on any required toolchain can instead be rejected regardless of their weighted score.

Acceptance and longer-but-cheaper proofs. Initialize $P^\star$ to the input proof. A candidate P' replaces $P^\star$ precisely when

$$\text{Verified}(P') \ \wedge \ \widehat{F}_{\mathbf{w}}(P') < \widehat{F}_{\mathbf{w}}(P^\star). \tag{4}$$

Thus, the current proof is the best-scoring verified proof encountered so far. The default weights $(1, 0, 0)$ recover the length-only acceptance shown in Algorithm 1, keeping the main comparisons aligned with prior length-optimizing work. Compilation-aware retrieval and this acceptance rule serve distinct roles: retrieval proposes relevant strategies, while acceptance determines which verified rewrites are retained. With $(0, 1, 0)$, a candidate with lower compilation cost is accepted regardless of its length, including when it is longer than the input. With both $w_1, w_2 > 0$ and $w_3 = 0$, an increase in length is accepted if its weighted cost is outweighed by the compilation-cost saving.

Weight-controlled evaluation. Table 7 compares length-only and heartbeat-only acceptance for Lean Refactor w/ Reranking on PutnamBench [44], using heartbeats as f_2 and fixing $w_3 = 0$. Both columns report mean per-proof relative reductions from the original proofs, following Appendix K.2.

Table 7: Effect of the acceptance objective for Lean Refactor w/ Reranking on PutnamBench. Larger reductions are better. Bold indicates the best result among the two settings for each metric. Heartbeats measure compilation work rather than wall-clock time.

Weights (w_1, w_2)	**Length reduction (%)**	**Heartbeat reduction (%)**
$(1, 0)$: length only	**64.29**	48.92
$(0, 1)$: heartbeats only	53.78	**54.41**

Changing the acceptance objective trades 10.51 percentage points of length reduction for 5.49 points of additional heartbeat reduction. The two aggregate outcomes are non-dominating relative to each other, demonstrating an observed trade-off rather than establishing global Pareto optimality.

Constrained variants. To minimize compilation cost under a length limit L, start with a verified proof of at most L tokens and accept only cheaper proofs that remain within the limit. Similarly, proof length can be minimized under a fixed compilation-cost limit.

G Retrieval Model Training Details

We initialize our strategy search model using Qwen3-Embedding-8B [51]. We fine-tune the model with a contrastive learning objective, using in-batch negative sampling and a margin-based false-negative filtering mechanism. We train the model for 1 epoch, a per-device batch size of 16, gradient accumulation steps of 2, and 4 NVIDIA 6000Ada GPUs. We use the AdamW optimizer with a learning rate of 6×10^{-6} and $\epsilon = 1 \times 10^{-8}$. The learning rate is linearly warmed up over the first 300 steps, then decayed to 0 using a cosine schedule. Temperature τ is set to 0.01, and the margin m is set to 0.1.

H Proof Dependency Reduction

An essential measure of formal proof elegance and maintainability is a minimal reliance on external lemmas. Analyzing the length-optimized proofs from Section 4.1, we find that Lean Refactor consistently reduces dependencies, defined as statements invoked from external libraries (e.g., Mathlib) or distinct intra-project modules. As shown in Tables 8 and 9, dependencies decreased in 64.0% of PutnamBench and 51.0% of miniF2F proofs. Remarkably, this trend holds even in complex research-level mathematics typically entangled with extensive library architectures, yielding reductions in 66.7% of the Analysis and 60.0% of the FLT datasets. Across all benchmarks, cases where the agent increases dependencies remain rare.

This systematic reduction reveals that Lean Refactor achieves brevity through genuine logical simplification rather than merely offloading complexity to heavy, specialized library calls. By discovering direct pathways from local premises, the agent naturally produces highly self-contained proofs. This structural shift provides a significant practical advantage by eliminating the burden of memorizing or invoking vast numbers of external lemma names. Consequently, these optimized proofs are significantly more readable for human mathematicians, easier to maintain, and inherently more resilient to the cross-version library breakages discussed in Section 4.3.

Table 8: Comparison of Proof Dependency Counts across miniF2F, PutnamBench, and Putnam 2025 datasets. Reduced Dependencies indicate a reduction in the number of external (e.g., Mathlib) or intra-project dependencies required for the proof.

Comparison Metric	miniF2F [53]	PutnamBench [44]	Putnam 2025
Total Theorems Compared	194	75	66
Reduced Dependencies	99 (51.0%)	48 (64.0%)	29 (43.9%)
Increased Dependencies	11 (5.7%)	11 (14.7%)	12 (18.2%)
No Change	84 (43.3%)	16 (21.3%)	25 (37.9%)

Table 9: Comparison of Proof Dependency Counts across Analysis, FLT, PFR, and Physlean datasets. Reduced Dependencies indicate a reduction in the number of external (e.g., Mathlib) or intra-project dependencies required for the proof.

Comparison Metric	Analysis [39]	FLT [9]	PFR [40]	Physlean [42]
Total Theorems Compared	45	45	45	45
Reduced Dependencies	30 (66.7%)	27 (60.0%)	24 (53.3%)	19 (42.2%)
Increased Dependencies	8 (17.7%)	11 (24.4%)	7 (15.6%)	11 (24.5%)
No Change	7 (15.6%)	7 (15.6%)	14 (31.1%)	15 (33.3%)

I Zero-shot Version Transfer Details

We complement the retrieval-side analysis with a passive audit of zero-shot version transfer: we take the proofs produced on a single source version (v4.19.0) and compile them against every subsequent Mathlib release through v4.29.0, recording how many still type-check without further refactoring. We compare our system (LeanRefactor) against the original human-written proofs and against ProofOptimizer, a shortening baseline run from the same source version. This probes whether the act of refactoring itself alters how well a proof transfers across Lean/Mathlib releases, independent of any retrieval-time filtering. Table 10 reports the results.

Table 10: Zero-shot version transfer of refactored proofs. Number of v4.19.0 proofs from each system that still type-check at each later Mathlib release without further refactoring. Totals: 75 (PutnamBench), 194 (MiniF2F).

Mathlib Version	PutnamBench			MiniF2F		
	LeanRefactor	ProofOptimizer	Original	LeanRefactor	ProofOptimizer	Original
v4.20.1	71	70	71	194	194	194
v4.21.0	69	70	**71**	193	194	194
v4.22.0	64	63	64	**189**	188	187
v4.23.0	51	50	**61**	**175**	170	151
v4.24.0	50	50	**60**	**172**	167	148
v4.25.0	49	48	**58**	**168**	162	141
v4.26.0	48	48	**58**	**168**	162	141
v4.27.0	45	45	**53**	**167**	162	140
v4.28.0	45	45	**53**	**166**	161	140
v4.29.0	43	42	**52**	**166**	161	140

The comparison against original long proofs is benchmark-dependent: on MiniF2F, LeanRefactor transfers furthest among all baselines, whereas on PutnamBench it falls short of the originals and tracks ProofOptimizer closely. We therefore conclude that whether shortening substitutes tactics that transfer better or worse across releases depends on corpus composition.

J Proof Dependencies in Research Repositories and Competition Problems

Table 11: Competition-type Lean sources rely on much fewer dependencies (Mathlib and intra-project) than Research-type Lean sources.

Type	Lean Source	Average Number of Dependencies Used
Competition	miniF2F [53]	2.66
	PutnamBench [44]	10.74
Research	FLT [9]	39.07
	PFR [40]	40.64
	PhysLean [42]	38.53
	Analysis [39]	35.38

K Experimental Settings

This section provides full details for the experiments summarized in Section 4. We describe (i) the seven evaluation benchmarks and how their initial long proofs are obtained, (ii) the metrics and the rationale for the benchmarks each metric is computed on, (iii) the baseline configurations used for the controlled ablation and the Claude Code comparison, and (iv) the implementation and hyperparameter settings shared across all runs.

K.1 Benchmarks

Competition problems. We evaluate on three competition benchmarks.

- **miniF2F** [53] (194 theorems). Initial long proofs are taken directly from the public release of our primary baseline ProofOptimizer [18], ensuring identical inputs across both methods.
- **PutnamBench** [44] (75 theorems). Initial long proofs are likewise taken from the ProofOptimizer release for direct comparability.
- **Putnam2025** (66 theorems). We construct this benchmark from the official Lean 4 solutions to problems B1 and B2 of the 2025 Putnam Competition produced by Seed-Prover 1.5 [10]. We restrict attention to B1/B2 due to budget. Each official solution is provided as a self-contained Lean 4 file; we extract the individual theorems and track the dependencies invoked in the original proof, yielding 66 theorems.

The length distribution of the initial proofs across all three competition benchmarks is shown in Figure 6.

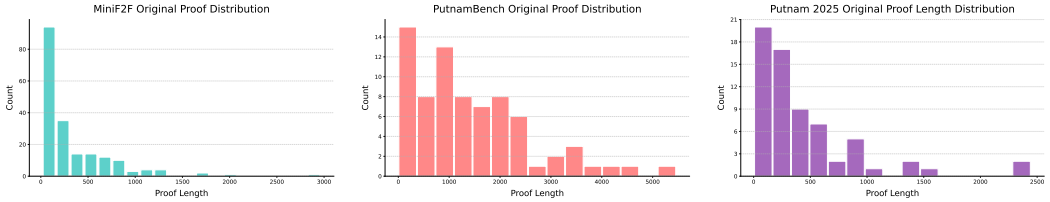


Figure 6: Distribution of original proof lengths across three competition style evaluation sets (miniF2F, PutnamBench, Putnam 2025).

Research-level problems. To assess refactoring on advanced mathematics, we draw theorems from four research-level Lean 4 repositories: Analysis [39], Fermat's Last Theorem (FLT) [9], the Polynomial Freiman–Ruzsa Conjecture (PFR) [40], and PhysLean [42]. From each repository we sample 45 theorems, subject to compute budget. To evaluate refactoring effectiveness across a wide range of initial proof complexities, we stratify the sample by proof length:

- 15 proofs exceeding 1000 tokens,
- 15 proofs between 500 and 1000 tokens,
- 10 proofs between 100 and 500 tokens,
- 5 proofs under 100 tokens.

When a repository contains insufficient theorems in a given length bucket, the unfilled quota is reallocated to the nearest non-empty bucket. The extraction pipeline retains all internal and external dependencies, ensuring that the agent operates on a properly contextualized, dependency-aware proof environment. The length distribution of the initial proofs across the four research-level evaluation sets is shown in Figure 7.

K.2 Metrics

We report three metrics across our benchmarks.

Proof length. We measure proof brevity as the average relative percentage decrease in token count between the initial long proof and the refactored proof. To ensure consistency with prior work, we use the syntax-aware tokenizer released by ProofOptimizer [18]; code is provided in Appendix M. This metric is reported on all seven benchmarks.

Compilation cost. We assess computational efficiency using `lake env lean -profile`, which exposes per-declaration timing. We isolate proof execution time from library import overhead, since import time is dominated by caching effects unrelated to proof content. Each measurement is taken in a consistent environment to ensure stable, deterministic timings, and we report the mean over five runs on identical hardware. On miniF2F and PutnamBench we follow the per-proof setting as each proof is a standalone `.lean` file, compiling each refactored proof in isolation. Research repositories do not admit clean per-proof wall-clock attribution, so we instead measure heartbeats, Lean 4's deterministic proxy for elaboration work; see Appendix O for full details. This metric is reported on all seven benchmarks.

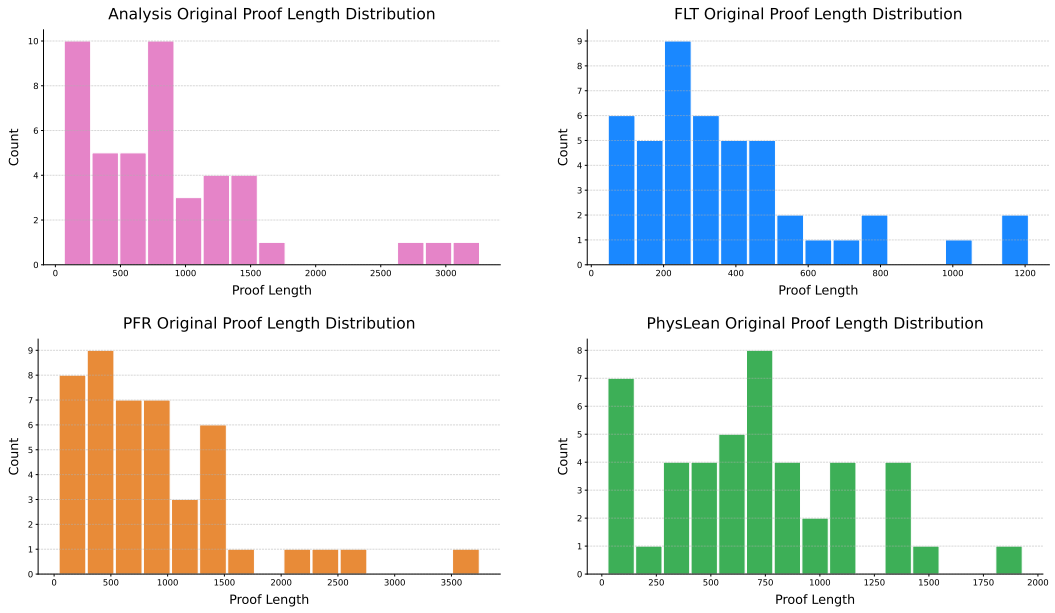


Figure 7: Distribution of original proof lengths across four research-level evaluation sets (Analysis, FLT, PFR, PhysLean).

Cross-version compatibility. To evaluate the impact of our version-tagged metadata, we compare the agent's performance under *version-filtered* strategy retrieval against an unfiltered baseline. Refactoring is run across three Lean and `Mathlib` toolchains (`v4.14.0`, `v4.16.0`, and `v4.22.0`); we then measure proof length reduction in each environment to determine whether retrieval tailored to a target version yields larger gains than version-agnostic retrieval. We additionally assess zero-shot version transfer by recompiling `v4.19.0` PutnamBench outputs from Lean Refactor, ProofOptimizer, and the original verbose proofs against later `Mathlib` releases through `v4.29.0` without further refactoring, measuring how many proofs continue to type-check across future toolchains.

K.3 Baselines

ProofOptimizer. On miniF2F and PutnamBench, we benchmark against ProofOptimizer [18] using their officially released shortest proofs, obtained via an 8-iteration shortening process with sampling budget 64 in the first six iterations and 1024 in the final two (denoted $6{\times}64 + 2{\times}1024$). Because the ProofOptimizer model checkpoints are closed-source, we cannot run their model on the remaining benchmarks (Putnam2025, Analysis, FLT, PFR, PhysLean). On all benchmarks, we conduct a controlled ablation over variants of our own framework.

Internal ablations. We define four internal baselines, each ablating a different component of Lean Refactor:

- **Base Agent.** The framework stripped to its core execution loop. The Planner module is removed; only the Refactorer and the verifier-guided Debugger remain. Strategy retrieval is applied as in our full method (Section 3.3), and the retrieved strategies are passed directly to the Refactorer. *This isolates the contribution of the Planner.*
- **Base Agent w/o Retrieval.** The Base Agent, with strategy retrieval (Section 3.2) additionally disabled. *This isolates the contribution of retrieval within the simplified loop.*
- **Lean Refactor w/o Retrieval.** The complete agent architecture (Planner, Refactorer, and Debugger) with the retrieval module disabled. *This isolates the gains attributable purely to the Planner's task decomposition, in the absence of external strategy context.*
- **Lean Refactor w/ Random Retrieval.** The full agent framework, but with our semantic and metadata-aware retrieval replaced by uniform random sampling from the strategy bank. This serves

as a *control for the embedding-and-rerank pipeline itself*, distinguishing the value of *retrieving the right* strategies from the value of *having any* strategies in context.

LLM-agnosticism and Claude Code comparison. We additionally run our framework with Claude Haiku 4.5 [5] to test that gains are not specific to Gemini 3 Flash. Using the same Claude Haiku 4.5 backbone, we compare against **Claude Code**. We prompt Claude Code with the same proof-shortening objective as our Lean Refactor agent, but adapted to its native abstractions (skills and MCP servers) rather than forcing it into our planner–refactorer–debugger loop. This lets Claude Code exercise its own agentic control flow on the task our framework is designed for, isolating the contribution of our architecture from the contribution of simply prompting a capable general-purpose agent to shorten Lean proofs. We evaluate Claude Code under three tool configurations of increasing support:

- **Lean MCP only.** The agent is provided only with the Lean MCP server[6], which exposes compiler feedback for correctness checking but no refactoring guidance.
- **Lean MCP + Lean skills.** The previous configuration is augmented with the Lean skills package[7], which supplies general Lean 4 refactoring and golfing guidance.
- **Lean MCP + Lean skills + our strategy retrieval tool.** The previous configuration is further augmented with a retrieval tool that exposes our strategy bank as an on-demand resource the agent can query during refactoring.

This comparison isolates the marginal value contributed by our strategy bank when added to a strong off-the-shelf coding agent. Due to compute budget, we run the Claude Code comparison on one representative benchmark per regime: PutnamBench (competition) and Analysis (research-level).

K.4 Implementation Details

Lean environment. All data collection and all experiments use Lean `v4.24.0` as the default environment, except where the cross-version compatibility study explicitly requires `v4.14.0`, `v4.16.0`, or `v4.22.0`.

Model serving. All open-weight models used during data collection (in particular GPT-OSS and the Qwen3-Embedding-8B) are primarily served via vLLM with default sampling parameters. Additionally, we utilized the OpenRouter API for GPT-OSS. For all GPT-OSS deployments, we configured the model using the medium reasoning-strength setting. Total compute for the reported experiments was approximately 10,000 GPU-hours on NVIDIA 6000 Ada GPUs, with strategy-bank construction (proof synthesis, strategy distillation, deduplication) accounting for the majority of this cost. API-served models (GPT-OSS-120B via OpenRouter, Gemini 3 Flash, and Claude Haiku 4.5) incurred an additional cost of approximately $5,000 USD.

Agent configuration. For our agent framework, we use:

- a maximum of 3 debugging cycles per refactoring episode ($D = 3$),
- a total API call budget of 30 per theorem ($B = 30$),
- the minimum proof length during optimization is 5 ($T = 5$),
- proof chunking with chunk sizes of 5, 10, and 20 lines. The multi-granularity chunking ensures that retrieval queries target proof regions of different lengths and consequently surface strategies appropriate to refactoring at different granularities.
- experiments were run on servers equipped with Intel Xeon w9-3495X CPUs and NVIDIA 6000 Ada GPUs.

Backbone LLMs. Unless stated otherwise, all main-paper experiments use Gemini 3 Flash [41] as the agent backbone. Claude Haiku 4.5 [5] is used both for the LLM-agnosticism check and as the shared backbone in the Claude Code comparison.

[6] `https://github.com/oOo0oOo/lean-lsp-mcp`
[7] `https://github.com/cameronfreer/lean4-skills`

L Run-to-Run Variance

To quantify variation from stochastic LLM generation, we repeated two representative comparisons three times under the same backbone, API budget, and evaluation protocol as the corresponding main-paper experiments. Running three independent repetitions for every method, benchmark, backbone, and Lean version was prohibitively costly. We therefore selected settings that represent the paper's two central retrieval claims: PutnamBench and FLT cover the competition and research regimes, respectively, for the targeted-versus-random retrieval comparison, while PutnamBench at `v4.22.0`, the latest evaluated target version, probes the compatibility-filtering mechanism. These experiments are intended to test whether the key gaps exceed run-to-run noise, rather than to provide exhaustive variance estimates for every cell in the main tables. Each entry below is the average relative proof-length reduction over the corresponding full benchmark; standard deviations are computed across the three independent runs.

Table 12: Run-to-run variance for targeted versus random retrieval. Average relative proof-length reduction (%) with Gemini 3 Flash and an API call budget of 30. Mean ± standard deviation is computed over three independent runs.

Benchmark	Method	Run 1	Run 2	Run 3	Mean ± std
PutnamBench	Lean Refactor	70.38	69.36	69.74	$\mathbf{69.83 \pm 0.52}$
	Lean Refactor w/ Random Retrieval	59.41	60.13	59.17	59.57 ± 0.50
FLT	Lean Refactor	23.55	23.31	23.12	$\mathbf{23.33 \pm 0.22}$
	Lean Refactor w/ Random Retrieval	21.47	20.99	21.28	21.25 ± 0.24

Targeted retrieval is clearly separated from run-to-run variation in both evaluation regimes. Its mean advantage over random retrieval is 10.26 percentage points on PutnamBench and 2.08 points on FLT, approximately 20 and 9 times the respective within-setting standard-deviation scales. The run ranges also do not overlap: the weakest Lean Refactor run exceeds the strongest random-retrieval run on both benchmarks. Thus, the gains arise from targeted strategy selection rather than from random strategy-bank context or run-to-run noise.

Table 13: Run-to-run variance for cross-version compatibility. Average relative proof-length reduction (%) on PutnamBench at Lean `v4.22.0`, using Gemini 3 Flash and an API call budget of 30. Mean ± standard deviation is computed over three independent runs.

Retrieval setting	Run 1	Run 2	Run 3	Mean ± std
Filtered Retrieval	57.36	57.83	57.85	$\mathbf{57.68 \pm 0.28}$
Full Retrieval	55.43	55.89	55.58	55.63 ± 0.23
No Retrieval	55.99	55.97	56.18	56.05 ± 0.12

Filtered Retrieval is likewise clearly separated from run-to-run variation. Its mean length reduction exceeds No Retrieval by 1.63 percentage points and Full Retrieval by 2.05 points, while the within-setting standard deviations are only 0.12–0.28; the filtered run range does not overlap either baseline. In contrast, Full Retrieval and No Retrieval remain close to the observed noise scale.

M Code for Getting Proof Length

```python
def proof_length(statement_and_proof):
    """
    Compute the token count of a proof from a full statement string.
    Extracts the proof by finding where the signature ends, then tokenizes
        and counts.
    """
    lean_operators = [
        ":=", "!=", "&&", "-.", "->", "←", "..", "...", "::", ":>",
        "<;>", ";;", "==", "||", "=>", "<=", ">=", "⁻¹", "?_"
    ]
```

```
    lean_operators_spaced = [" ".join(conn) for conn in lean_operators]
    lean_operators_dict = dict(zip(lean_operators_spaced, lean_operators,
        strict=False))

    def lexer(lean_snippet):
        tokenized_lines = []
        for line in lean_snippet.splitlines():
            tokens = []
            token = ""
            for ch in line:
                if ch == " ":
                    if token:
                        tokens.append(token)
                        token = ""
                elif str.isalnum(ch) or (ch in "._'"):
                    token += ch
                else:
                    if token:
                        tokens.append(token)
                        token = ""
                    tokens.append(ch)
            if token:
                tokens.append(token)
            tokenized_line = " ".join(tokens)
            for conn in lean_operators_spaced:
                if conn in tokenized_line:
                    tokenized_line = tokenized_line.replace(conn,
                        lean_operators_dict[conn])
            tokenized_lines.append(tokenized_line)
        return "\n".join(tokenized_lines)

    try:
        statement_and_proof = remove_comments(statement_and_proof) # Remove
            all comments in the proof
        decl_start, decl_end = return_theorem_to_prove(statement_and_proof) #
             Gets the declaration start and end
        proof = statement_and_proof[decl_end:] # Isolate the proof from the
            declaration
        proof_tokenized = lexer(proof)
        return sum([len(l.split(" ")) for l in proof_tokenized.splitlines()])
    except:
        return 10**9
```

N Comparison with ProofOptimizer's Heartbeat-Optimized Proofs

The wall-clock results in Table 4 compare against ProofOptimizer's *length-optimized* proofs because their heartbeat-optimized model weights are not released, putting their best-case compilation numbers out of reach of direct re-execution. We close this gap using Lean's `#count_heartbeats` (under `set_option Elab.async false`) as a deterministic proxy of compilation cost, the same metric ProofOptimizer adopts when optimizing for execution efficiency. This permits a direct comparison against the compilation cost numbers that ProofOptimizer [18] publish: the values produced by their heartbeat-optimized inference-time variant after eight iterations with sampling budget $6\times64 + 2\times1024$, i.e. their strongest reported heartbeat results. We compute the heartbeats for proofs refactored under the *Lean Refactor w/ Reranking* setting from Section 4.2, and report the comparison in Table 14.

Objective-conditioned reranking yields the lowest heartbeat counts. Across both benchmarks, Lean Refactor w/ Reranking attains the lowest average heartbeat count among all evaluated methods, improving over ProofOptimizer's heartbeat-optimized proofs from 10.4K to 9.2K on miniF2F and from 111K to 82K on PutnamBench. We emphasize that ProofOptimizer's underlying model is trained to generate shorter proofs; its heartbeat-optimized variant achieves the published numbers by reusing the same length-trained model and selecting low-heartbeat candidates at inference. This exposes a train–inference objective mismatch: the prior is shaped by token length, while the selection rule targets execution cost. Our framework avoids this mismatch by carrying the objective in the

Table 14: Heartbeat-level comparison against ProofOptimizer's heartbeat-optimized variant. Heartbeats reports the average heartbeat count (in thousands) across theorems. ProofOptimizer values are taken from their heartbeat-optimized proofs; "Original" refers to the unrefactored proofs from the same source. Lower is better; best per column in bold.

Method	miniF2F Heartbeats (K) ↓	PutnamBench Heartbeats (K) ↓
Original (no refactoring)	36.3	221
ProofOptimizer, Heartbeat-optimized [18]	10.4	111
Lean Refactor w/ Reranking (Ours)	**9.2**	**82**

retrieval rule rather than in model weights, reranking a shared, semantically-retrieved candidate pool by per-strategy heartbeat metadata (Section 3.2). Because the metadata is annotated directly on strategies, the same bank can be redirected to new cost metrics without retraining.

O Compilation Cost for Research-Level Proofs

For competition benchmarks, each proof lives in a self-contained `.lean` file, so we can isolate proof execution from import overhead and report wall-clock time directly. Research repositories invert this setup: proofs reside inside large multi-file Lean projects whose elaboration is interleaved with import resolution, neighboring declarations, and kernel cache state, and there is no stable way to attribute wall-clock time to a single proof in this regime[8]. We therefore switch to *heartbeats*, Lean 4's deterministic proxy for elaboration work. We prepend each proof with `set_option Elab.async false in` and `#count_heartbeats in` to pin the count to the declaration under measurement, consistent with prior work [18]. We report dataset-level means of the per-proof absolute and relative reductions in Table 15.

Table 15: Heartbeat reduction on Analysis [39], FLT [9], pfr [40], and PhysLean [42]. *Avg. Rel. Reduc.*: average per-proof relative % decrease in heartbeats. *Avg. Abs. Reduc.*: average per-proof absolute decrease in heartbeats. Higher is better for both. Each dataset contains $n = 45$ proofs.

Dataset	Method	Avg. Rel. Reduc. ↑ (%)	Avg. Abs. Reduc. ↑ (heartbeats)
Analysis [39]	Lean Refactor w/o Retrieval	−6.72	2324
	Lean Refactor w/ Random Retrieval	−4.18	2759
	Lean Refactor	−14.43	3049
	Lean Refactor w/ Reranking	**3.19**	**3809**
FLT [9]	Lean Refactor w/o Retrieval	0.37	−946
	Lean Refactor w/ Random Retrieval	0.70	222
	Lean Refactor	3.00	311
	Lean Refactor w/ Reranking	**11.33**	**1235**
PFR [40]	Lean Refactor w/o Retrieval	−6.81	1332
	Lean Refactor w/ Random Retrieval	9.21	3223
	Lean Refactor	3.13	2184
	Lean Refactor w/ Reranking	**9.89**	**3309**
PhysLean [42]	Lean Refactor w/o Retrieval	1.41	1152
	Lean Refactor w/ Random Retrieval	3.73	2166
	Lean Refactor	-21.97	2609
	Lean Refactor w/ Reranking	**7.47**	**4101**

Lean Refactor with reranking is the strongest method by both metrics on every dataset, indicating that objective-aligned retrieval, not structural rewriting alone, is what consistently delivers heartbeat savings. The two columns can disagree in sign (notably for Analysis, PFR, and PhysLean under non-reranking variants) because the relative mean is sensitive to outliers: a few refactored proofs exhibit large heartbeat increases that pull the average relative reduction negative, even when the absolute mean remains positive and the majority of proofs improve.

[8] `https://github.com/leanprover/lean4/issues/8038`

P Additional Proof Length Reduction Results

We further evaluate Lean Refactor on two additional benchmarks from distinct domains: **Verina** [50], a benchmark for verifiable code generation in Lean, from which we use the proof-generation task, and a **PDE** [38] benchmark of formalized statements about partial differential equations.

Verina setup. Verina consists of 189 problems. Since no public set of model-generated proofs is available, we first obtain solutions with a refinement-loop theorem prover built on top of Claude Opus 4.6: for each task the model generates a candidate proof, the Lean compiler checks it, and compiler errors are fed back as additional context, with up to 16 attempts at temperature 1.0 (matching the Verina default). Each candidate is validated with integrity checks (no modification to the original code, specification, or theorem statement) and axiom checks (only `propext`, `Classical.choice`, and `Quot.sound` are permitted). This solver closes 14/189 problems (7.4%; 1 advanced and 13 basic tasks), yielding 15 theorems to which we then apply Lean Refactor.

PDE setup. We follow the same sampling strategy specified in Appendix K.1 and sample 45 theorems spanning diverse proof lengths.

Results. Lean Refactor framework achieves an average relative proof-length reduction of **53.50%** on Verina and **35.34%** on PDE. This indicates that the framework transfers cleanly from competition mathematics to verifiable code generation (Verina [50]) and to formalized PDE topics (PDE [38]), supporting the claim that Lean Refactor's refactoring performance reflects a general capability for optimizing Lean proofs across mathematical and verification domains.

Q Examples of Refactoring Strategies

Strategy 1: Eliminate exhaustive numeric case analysis; construct new witness via arithmetic shift

Description. A wasteful proof pattern: extract concrete numeric values for an existential witness (and any auxiliary constants) by evaluating a universal hypothesis at many points and discharging the resulting system with `norm_num`, `nlinarith`, and `linarith`. This is unnecessary when the goal differs from the hypothesis only by a shift in the bound variable — the witness can be obtained by translating the existing one algebraically, leaving the numeric values abstract.

When to apply. The hypothesis has the form `∃ x, ∀ y, P x y` and the goal has the form `∃ x, ∀ y, P x (y - k)` (or any other affine reparametrisation of `y`). **Note:** The remaining goal after providing the shifted witness must close by specializing the hypothesis at the shifted point — i.e. it must not depend on the specific numeric value of the witness. If it does, the original numeric derivation is needed.

Application guide.

1. Delete every block that pins the witness or auxiliary constants to specific numeric values — the long chains of `have h_i := hyp v_i` followed by `norm_num`, `nlinarith`, `linarith`.
2. Destructure the existential: `rcases h with ⟨a, ha⟩`.
3. Provide the shifted witness: `refine ⟨a + k, ?_⟩`, where k is the shift between the hypothesis's bound variable and the goal's (read off the goal: if it mentions `y - k`, use `a + k`).
4. Close the remaining goal by specializing `ha` at the shifted point, typically `intro y; simpa using ha (y - k)`, or a short `linarith`/`ring` after substitution.

Example. Schematic shape: from a hypothesis `h : ∃ x, ∀ y, P x y` prove the shifted goal `∃ x, ∀ y, P x (y - k)`. The before-proof wastefully derives the concrete value of the witness a (and a secondary constant c) by instantiating `h` at many integers.

Before

```
-- pin the witness to a concrete value
have h_a : a = a0 := by
  have e1 := h v1
  have e2 := h v2
  -- ... many more instantiations ...
  have e_n := h v_n
  norm_num at e1 e2 ... e_n
  nlinarith
-- pin an auxiliary constant
have h_c : c = c0 := by
  have e1 := h v1_prime
  -- ... many more instantiations ...
  have e_m := h v_m_prime
  norm_num [h_a] at e1 ... e_m |-
  <;> linarith
use a0 + k
-- ... finish using h_a, h_c ...
```

After

```
rcases h with <a, ha>
refine <a + k, ?_>
intro y
simpa using ha (y - k)
```

Reduction: large.

Strategy 2: Replace per-entry `have` lemmas with a single `ext` and `fin_cases` proof

Description. The long proof proves each matrix entry separately using `have` statements and then glues them together with `Matrix.ext` inside a final `h_main`. For matrices indexed by a finite type, equality can be reduced to pointwise equality by `ext`, after which a case split on the indices (`fin_cases`) makes each entry goal trivial. A single `simp` (with the right lemmas) then closes all cases, removing the need for separate per-entry lemmas.

When to apply. Use this pattern whenever you are proving equality of two matrices (or similar objects) indexed by `Fin n` or another finite type, and you have written individual lemmas for each entry. **Note:** Requires that the target matrix is built via `Matrix.of` / `!![...]` notation (so `Matrix.of_apply` reduces it) and that the per-entry goals reduce to something a unified finisher can solve. If the original per-entry proofs required nontrivial work, that work must still be expressible as part of the unified finisher.

Application guide.

1. Delete all `have` statements that compute each entry individually, including the final `h_main` that wraps `apply Matrix.ext`.
2. After `intro X`, write `ext i j`.
3. Immediately follow with `fin_cases i <;> fin_cases j` to enumerate all index pairs.
4. Finish with

   `simp [Matrix.mul_apply, Matrix.of_apply, Fin.sum_univ_two]`

   optionally followed by `<;> ring` or `<;> norm_num` if arithmetic remains.
5. Remove the final `exact h_main` as the goal is already solved.

Example. *Before*

```
intro X
have h1 : ((M : Matrix (Fin 2) (Fin 2) R) * X) 0 0 = a := by
```

```
  simp [Matrix.mul_apply]
  -- many try tactics
have h2 : ((M : Matrix (Fin 2) (Fin 2) R) * X) 0 1 = b := by
  simp [Matrix.mul_apply]
  -- many try tactics
have h3 : ((M : Matrix (Fin 2) (Fin 2) R) * X) 1 0 = c := by
  simp [Matrix.mul_apply]
  -- many try tactics
have h4 : ((M : Matrix (Fin 2) (Fin 2) R) * X) 1 1 = d := by
  simp [Matrix.mul_apply]
  -- many try tactics
have h_main : (M : Matrix (Fin 2) (Fin 2) R) * X = M2 := by
  apply Matrix.ext
  intro i j
  fin_cases i <;> fin_cases j
  simp [h1, h2, h3, h4]
exact h_main
```

After

```
intro X
ext i j
fin_cases i <;> fin_cases j <;>
  simp [Matrix.mul_apply, Matrix.of_apply, Fin.sum_univ_two]
```

Reduction: medium.

Algorithm 1 Retrieval-Augmented Agent Framework for Proof Refactoring

Require: Initial unoptimized proof P, Strategy DB $\mathcal{S}$, Budget B, Target length threshold T, Max debug rounds D, Retrieval budget K
Ensure: Optimized refactored proof P^*

$P_{curr} \leftarrow P$
$History \leftarrow \emptyset$
$steps \leftarrow 0$
while $steps < B$ **and** $\textsc{Length}(P_{curr}) > T$ **do**
  $Chunks \leftarrow \textsc{SegmentProof}(P_{curr})$
  $\mathcal{C}_{retrieved} \leftarrow \emptyset$
  for $c \in Chunks$ **do**
    $E_c \leftarrow \textsc{Embed}(c)$
    $S_c \leftarrow \textsc{RetrieveAndRerank}(\mathcal{S}, E_c, K)$ ▷ Multi-objective alignment
    $\mathcal{C}_{retrieved} \leftarrow \mathcal{C}_{retrieved} \cup \{(S_c, \textsc{Location}(c))\}$
  end for
  $Plan \leftarrow \textsc{Planner}(P_{curr}, \mathcal{C}_{retrieved}, History)$
  if $Plan = \emptyset$ **then**
    break ▷ Terminate if no refactoring steps are viable
  end if
  $step_applied \leftarrow$ **False**
  for $step \in Plan$ **do**
    $P_{cand} \leftarrow \textsc{Refactorer}(P_{curr}, step)$
    $success, \epsilon \leftarrow \textsc{CompileLean}(P_{cand})$
    $d \leftarrow 0$
    ▷ Verifier-guided self-debugging loop
    while not $success$ **and** $d < D$ **do**
      $P_{cand} \leftarrow \textsc{Debugger}(P_{cand}, \epsilon, step)$
      $success, \epsilon \leftarrow \textsc{CompileLean}(P_{cand})$
      $d \leftarrow d + 1$
    end while
    ▷ Check for successful compilation and length reduction
    if $success$ **and** $\textsc{Length}(P_{cand}) < \textsc{Length}(P_{curr})$ **then**
      $P_{curr} \leftarrow P_{cand}$
      $History \leftarrow History \cup \{(step, \text{Success})\}$
      $step_applied \leftarrow$ **True**
      break ▷ Break inner loop to dynamically replan with updated proof
    end if
  end for
  if not $step_applied$ **then**
    $History \leftarrow History \cup \{(Plan, \text{Failed})\}$
  end if
  $steps \leftarrow steps + 1$
end while
return P_{curr}

R Prompts

R.1 Planner Prompt

```
Your task is to analyze a given Lean 4 proof and create a structured plan for
optimizing and shortening it, while maintaining the correctness.

## Instructions

Analyze the proof and identify regions (or entire proof) that can be optimized
and simplified. For each optimization opportunity:
```

````
1. **Identify the line range**: ...
2. **Provide a title**: ...
3. **Provide potential reduction**: ...
4. **Determine an optimization strategy**: ...

## Inputs

1. A correct Lean 4 statement and proof source code.
2. Elaborated signature.
3. Doc string from the source if it exists.
4. All the dependencies used in the statement and proof...
5. A list of retrieved optimization strategies from a strategy index...

## Retrieved Optimization Strategies
...

## Output Format

Output the plans using the following JSON format, wrapped in a single ```json
``` tags. Ensure the output is valid JSON.

[
  {
    "line_start": X,
    "line_end": Y,
    "title": "the strategy name",
    "reduction": "high, medium, or low",
    "description": "Detailed description of the optimization strategy ..."
  }
]

Rules about the ordering of the output plans:
1) Sort plans primarily from top to bottom of the proof (increasing line numbers
).
2) Overlaps of optimization regions are allowed. If two plans overlap in
location, put the plan with the MOST significant potential reduction first.
3) If two plans have the same overlap and same potential impact, keep top-to-
bottom ordering.

Examples:
...

Here is the proof to optimize:
...

Theorem's elaborated signature:
...

Theorem's documentation:
...

Information about dependencies used in the proof:
...

Now analyze the proof and generate your optimization plans.
````
````

R.2 Refactor Prompt

```
You are an expert Lean 4 proof optimizer. Your task is to execute a specific
optimization plan on a proof, modifying the targeted section while ensuring the
proof remains correct.

## Current Proof
...

## Theorem's Elaborated Signature
...

## Documentation
...

## Dependencies
...

## Optimization Plan

**Target Lines:**: ...
**Title:**: ...
**Description:**: ...

## Instructions

1. **Focus on the targeted section**: ...
2. **Preserve correctness**: ...
3. **Do NOT modify the theorem statement**: ...
4. **Aim for shorter proof**: ...
5. **Output the complete theorem and proof**: ...

Before providing the code, briefly explain what changes you're making and why.

Then output the complete optimized Lean 4 code wrapped in tags:

‘‘‘lean4
<your optimized complete theorem and proof here>
‘‘‘
```

R.3 Debugger Prompt

```
The proof (Round {{ prev_round_num }}) is not correct. Following is the
compilation error message, where we use <error></error> to signal the position
of the error.

Error message:
...

Before producing the Lean 4 code that fixes the error, provide a detailed
analysis of the error message. Focus on fixing the compilation errors, and try
your best to preserve the proof optimization that was previously done. Do not
revert to the original unoptimized proof.

You must wrap the entire Lean 4 theorem and proof in tags like:

‘‘‘lean4
<your corrected proof here>
‘‘‘
```

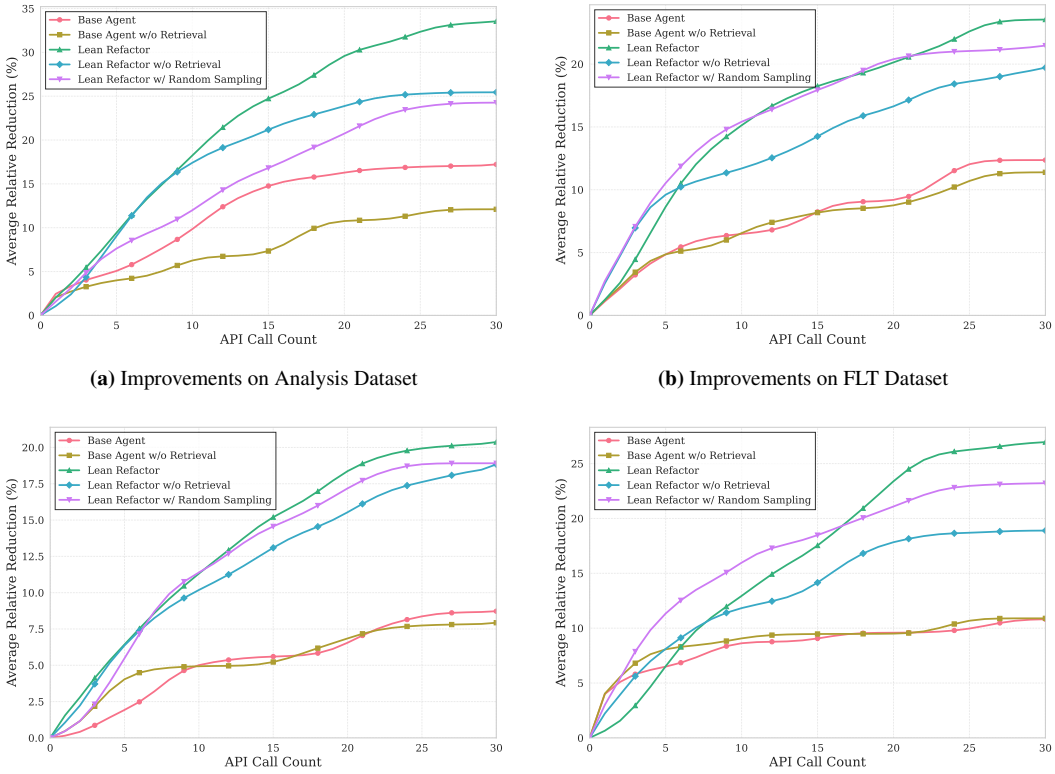
S Licenses for Existing Assets

All existing assets used in this paper are listed in Table 16 along with their licenses and source URLs. All assets are used in accordance with their respective terms of use.

Table 16: Licenses for existing assets used in this paper.

Asset	License	Source
Benchmarks		
miniF2F [53]	Apache 2.0	`https://github.com/openai/miniF2F`
PutnamBench [44]	Apache 2.0	`https://github.com/trishullab/PutnamBench`
Putnam2025 (Seed-Prover 1.5) [10]	Apache 2.0	`https://github.com/ByteDance-Seed/Seed-Prover`
Lean Repositories		
Mathlib4	Apache 2.0	`https://github.com/leanprover-community/mathlib4`
Analysis [39]	Apache 2.0	`https://github.com/teorth/analysis`
FLT [9]	Apache 2.0	`https://github.com/ImperialCollegeLondon/FLT`
PFR [40]	Apache 2.0	`https://github.com/teorth/pfr`
PhysLean [42]	Apache 2.0	`https://github.com/leanprover-community/physlib`
Datasets		
NuminaMath-1.5 [22]	Apache 2.0	`https://huggingface.co/datasets/AI-MO/NuminaMath-1.5`
FineLeanCorpus [34]	Apache 2.0	`https://huggingface.co/datasets/m-a-p/FineLeanCorpus`
ATLAS [25]	Apache 2.0	`https://github.com/Atlas-Lean/Atlas`
Models		
Qwen3-Embedding-8B [51]	Apache 2.0	`https://huggingface.co/Qwen/Qwen3-Embedding-8B`
Goedel-Prover-V2-32B [24]	Apache 2.0	`https://github.com/Goedel-LM/Goedel-Prover-V2`
GPT-OSS-120B / 20B [30]	Apache 2.0	`https://huggingface.co/openai/gpt-oss-120b`
Tools		
Lean MCP (lean-lsp-mcp)	MIT	`https://github.com/oOo0oOo/lean-lsp-mcp`
Lean 4 skills	MIT	`https://github.com/cameronfreer/lean4-skills`

T Test Time Scaling Curves



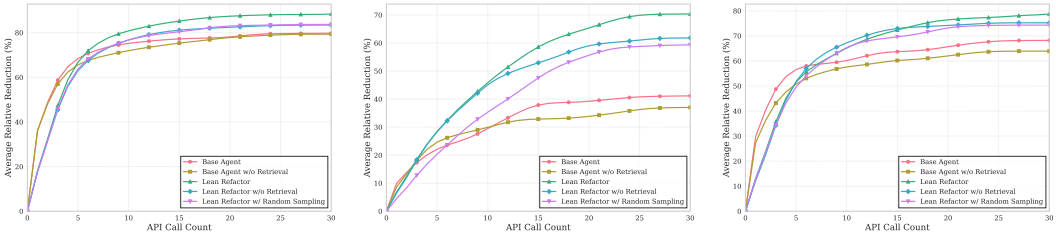
(a) Improvements on Analysis Dataset

(b) Improvements on FLT Dataset

(c) Improvements on PFR Datasets

(d) PhysLean Improvements

Figure 8: Average relative length reduction across research-level datasets over successive API calls. We compare the base agent against our proposed Lean Refactor agent, alongside ablation variants (without retrieval and with random sampling). The Lean Refactor agent converges to a higher relative reduction compared to all other variants.



(a) miniF2F Improvements

(b) PutnamBench Improvements

(c) Putnam 2025 Improvements

Figure 9: Average relative length reduction across competition-style datasets over successive API calls. We compare the base agent against our proposed Lean Refactor agent, alongside ablation variants (without retrieval and with random sampling). The Lean Refactor agent converges to a higher relative reduction compared to all other variants.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Every abstract claim maps to a specific result: over 70% compression on competition benchmarks (Table 1), and over 20% on research repositories (Table 2), and outperforming Claude Code (Table 3); up to 60% compilation-time reduction and outperforming ProofOptimizer (Tables 4); version-filtered retrieval gains (Table 5) and zero-shot version transfer (Table 6); and LLM-agnostic operation across Gemini, Claude, and GPT-OSS (Tables 3, 5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 and Appendix C discuss two limitations: (i) cross-version evaluation is restricted to PutnamBench because research-level projects are pinned to specific Lean/Mathlib toolchains, and (ii) strategy-level metadata uses conservative cluster-level aggregation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We release the implementation of our agentic system, and document the details of hyperparameters and experimental settings we used in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the implementation of our agentic system, and document the details of hyperparameters and experimental settings we used in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Benchmarks, metrics, baselines, and implementation details are in Appendix I. Retrieval-model training details are in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Compilation-time results in Table 4 report means with subscripted across-run standard deviations computed over five repeated compilations per proof on identical hardware.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix I.4 reports exact hardware, total compute, and API costs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Appendix O discusses positive impacts and notes no direct path to harmful applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released assets pose no foreseeable misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets (benchmarks, source repositories, datasets, models, tools) are cited in-text and listed with licenses and source URLs in Appendix P, Table 13. All are used in accordance with their public terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces 200K long–short proof pairs, 9K curated refactoring strategies, and a fine-tuned Qwen3-Embedding-8B retrieval model. The strategy schema (Section 3.1), construction pipeline (Appendix B), retrieval-training details (Appendix E) are all documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are core components of the methodology and their usage is documented throughout the paper: GPT-OSS-120B and Goedel-Prover-V2-32B for proof synthesis and strategy distillation, Gemini 3 Flash, Claude Haiku 4.5, and GPT-OSS-20B as the agent backbone for planning, refactoring, and debugging, and Qwen3-Embedding-8B as the strategy retrieval encoder.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.