SHADOWBENCH: Toward Reliable Automatic Evaluation of Semantic Alignment in Autoformalization

Hojae Han[1*] **Jongyoon Kim**[2*] **Sanghyeok Park**[2*] **Dongwook Cheon**[2] **Yeachan Park**[3]
Myung Jae Jeon[4] **Sunjong Choe**[2] **Soonho Kong**[5] **Wonseok Hur**[2]
Seung-won Hwang[2†] **Donghoon Hyeon**[2†]

[1]Electronics and Telecommunications Research Institute [2]Seoul National University
[3]Sejong University [4]University of Maryland, College Park [5]Amazon Web Services

hojae.han@etri.re.kr
{john.jongyoon.kim,202123018,seungwonh,dhyeon}@snu.ac.kr

Abstract

Autoformalization translates informal mathematical theorems into code for proof assistants such as Lean. A central challenge is that current evaluation metrics can accept type-correct but misaligned statements or reject correct statements written in a different formulation. Inspired by Pass@k, we propose SA-PASS (*Semantic Alignment Pass*), which tests formal statements using auxiliary statements called *shadows* that characterize the intended statement. A generated statement receives full credit only when it compiles, implies each shadow (forward check), and is implied by their conjunction (backward check). We instantiate SA-PASS in SHADOWBENCH, a Lean 4 full autoformalization benchmark of 178 postgraduate- to research-level problems spanning eight mathematical areas. Claude Code (Opus 4.8) with Numina-Lean-Agent reaches 61.8% compile rate and 11.2% SA-PASS. Across outputs generated by six agentic configurations, SA-PASS achieves 98.8% binary agreement with expert judgments. An early version of SHADOWBENCH served as the benchmark for Track 4 of the ICML 2026 AI4Math Challenge.[1]

1 Introduction

Autoformalization translates informal mathematics written in natural language into code for proof assistants such as Lean.[2] In the full autoformalization setting, the input is an informal theorem statement and proof, and the output is code containing the corresponding formal statement and proof. The proof assistant can then mechanically check whether the formal proof proves the formal statement.

The core challenge for evaluation is to assess whether the generated formal statement expresses the same mathematics as the informal statement (Ospanov et al., 2025; Ammanamanchi et al., 2026). Existing automatic metrics either rely on compile rate signals that can accept misaligned statements, assign scores without formally checking the intended meaning, or require equivalence to a fixed reference formulation that can reject correct statements expressed differently (§2.2). An alternative is manual judgment by mathematicians familiar with Lean, which is reliable yet too costly to scale for repeated evaluations (Wu et al., 2022; Fang et al., 2025; Ying et al., 2024).

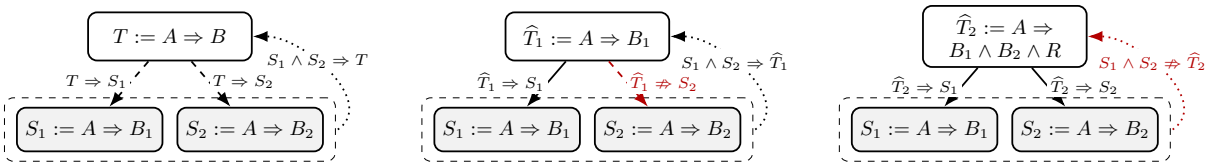
In this paper, we propose SA-PASS (*Semantic Alignment Pass*), an automatic metric that scores semantic alignment through forward and backward implication checks. SA-PASS follows the Pass@k idea from code generation, where an output is accepted only if it passes all tests for the task (Chen et al., 2021). For formal statements, we define these checks using a complete set of *shadow theorems*: auxiliary theorems selected such that each shadow is implied by the intended theorem and the shadows together characterize it (Figure 1a). During evaluation, a generated formalization receives the maximum SA-PASS score only if it compiles, its statement implies every shadow in the forward checks (Figure 1b), and its statement is implied by the shadows' conjunction in the backward check (Figure 1c).

This design makes expert supervision reusable: a shadow set is constructed once for each intended theorem rather than separately for every generated formalization. Experts construct the set by revising LLM-generated drafts (§4.1), and Lean mechanically verifies the set completeness (§3.2). The same shadow checkers can then be reused to evaluate any number of generated formalizations automatically.

Alongside SA-PASS, we introduce SHADOWBENCH, a Lean 4 full-autoformalization bench-

*These authors contributed equally to this work.
†Corresponding authors.
Contact information for all authors is provided in Appendix L.
[1]https://github.com/ldilab/shadowbench
[2]https://lean-lang.org/

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 53031–53067
October 24-29, 2026 ©2026 Association for Computational Linguistics



(a) Intended theorem and shadows. (b) Forward check failure. (c) Backward check failure.

Figure 1: Illustrative example of shadow theorem checks. Here A, B, B_1, B_2 are predicates and $B \Leftrightarrow B_1 \wedge B_2$. The intended theorem statement T implies both shadows (S_1, S_2), and the two shadows jointly imply T. The shadow theorem statement checks reject both generated theorems. The forward check finds that $\widehat{T}_1$ does not imply S_2, and the backward check finds that $S_1 \wedge S_2$ does not imply $\widehat{T}_2$.

mark comprising 178 postgraduate- to research-level problems across eight mathematical areas and three difficulty levels. Each problem is annotated with a complete set of shadow theorems, enabling automatic evaluation with SA-PASS. Existing benchmarks such as ProofNet (Azerbayev et al., 2023) primarily contain short formalizations with a single conclusion. In contrast, SHADOWBENCH targets multi-conclusion problems whose reference formalizations are about 1.6 times longer lines on average (Table 9). An early version of SHADOWBENCH, comprising 126 problems, served as the benchmark for Track 4 of the ICML 2026 AI4Math Challenge, with analyses of the results provided in Appendix G.

Empirically, Claude Code (Opus 4.8) with Numina-Lean-Agent (Liu et al., 2026a) achieves a 61.8% compile rate on SHADOWBENCH, yet only 11.2% of its outputs pass SA-PASS (Table 3). SA-PASS achieves 98.8% binary agreement with expert judgments for generated outputs across six agentic configurations, whereas compile rate and BEq+ (Liu et al., 2025a; Poiroux et al., 2025b) achieve 17.8% and 80.6%, respectively (Table 4).

We summarize our contributions as follows.

- We introduce SA-PASS, an automated metric that evaluates semantic alignment between formal and informal statements through Lean-checkable shadow theorems.
- We construct SHADOWBENCH, a Lean 4 full autoformalization benchmark that uses SA-PASS as its main evaluation metric across 178 postgraduate- to research-level problems. Claude Code (Opus 4.8) with Numina-Lean-Agent reaches 61.8% compile rate and 11.2% SA-PASS.
- SA-PASS achieves 98.8% binary agreement with expert judgments.

2 Related Work

2.1 Autoformalization Tasks and Benchmarks

Autoformalization is a task that translates an informal theorem into a formal theorem in a proof assistant such as Lean, for either the statement or the proof. For statement autoformalization, various methods have been proposed (Wu et al., 2022; Liu et al., 2025b; Lu et al., 2025b; Fang et al., 2025; Chen et al., 2026) and evaluated on benchmarks spanning difficulty levels from high school to undergraduate (Gao et al., 2025; Ying et al., 2024; Murphy et al., 2024; Xie et al., 2025). Lu et al. (2024); Jana et al. (2026) have studied full autoformalization, which produces both the statement and its proof from natural language. At the proof level, this task is mostly discussed as theorem proving, where various methods (Yang et al., 2023; Xin et al., 2024; Lin et al., 2025a; Wang et al., 2025; Liu et al., 2026a) generate a formal proof given the formal statement, evaluated on benchmarks such as miniF2F and ProofNet (Zheng et al., 2022; Azerbayev et al., 2023), and on research-level Lean projects (Poiroux et al., 2025a). While these benchmarks establish the task setting, our focus is the complementary problem of evaluating whether generated formal statements are semantically aligned with the intended theorem.

2.2 Semantic Alignment Evaluation

Compile rate. Compile rate runs the generated Lean code through the Lean compiler and counts it as correct when the generated theorem statement and proof type-check (Azerbayev et al., 2023). It rejects ill-formed or unproved Lean code, but it can produce false positives when a generated proof type-checks for a statement that differs from the informal theorem (Liu et al., 2025a; Ammanamanchi et al., 2026).

Reference-based metrics. Several metrics evaluate generated statements against a reference formalization. BLEU (Azerbayev et al., 2023; Papineni et al., 2002) measures lexical overlap with the reference theorem, TransTED (Liu et al., 2026b) compares structural similarity to the reference theorem, and FormalAlign (Lu et al., 2025a) and LLM-as-judge (Guo et al., 2025b) score the generated statement against the informal statement. Because these lexical, structural, and model-based scores are not implication checks, they can produce both false positives and false negatives. BEq and BEq+ (Liu et al., 2025a; Poiroux et al., 2025b) use Lean tactics to prove equivalence with the reference statement, but they create false negatives by rejecting correct statements whose formulations differ from the fixed reference (Table 4).

Expert judgments. Manual judgment by mathematicians familiar with Lean remains reliable, but it does not scale to large or repeated evaluations (Wu et al., 2022; Fang et al., 2025; Ying et al., 2024).

Test-based metrics. Testing Accuracy (Kim et al., 2026) treats the dependent successor theorems as test cases, accepting a generated theorem when all of them compile. This approach can produce false positives, since passing all successor theorem tests does not guarantee semantic alignment with the intended informal statement.

Our distinction. SA-PASS addresses both false positives and false negatives in semantic alignment by evaluating generated statements against a complete set of auxiliary theorems. SA-PASS verifies generated statement with a complete set of both forward and backward implication checks, obtaining 98.8% binary agreement with expert judgments (Table 4).

3 SA-PASS: Semantic Alignment Pass

3.1 Task and Shadow Theorems

Definition 1 (Full autoformalization task)**.** *Given an informal theorem statement and its proof, the full autoformalization task is to generate a formal statement $\widehat{T}$ together with a machine-checkable proof of $\widehat{T}$. Let T denote the intended formal statement. The semantic alignment objective is $\widehat{T} \Leftrightarrow T$.*

Definition 2 (Shadow theorem)**.** *For an intended formal statement T, a shadow theorem consists of an auxiliary formal statement S and a machine-checkable proof of S. Its forward checker theorem states that T implies S:*

$$T \Rightarrow S.$$

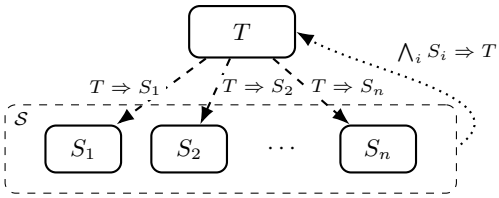


Figure 2: A complete shadow set. Each shadow statement S_i follows from the intended formal statement T. The set is complete when the shadows jointly imply T.

3.2 Complete Shadow Sets

Definition 3 (Complete shadow set)**.** *Let T be the intended formal statement, and let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a finite set of shadow statements. The set $\mathcal{S}$ is* complete *if the shadows jointly characterize T, i.e., the backward checker theorem states*

$$S_1 \wedge \cdots \wedge S_n \Rightarrow T.$$

Together with the forward checker theorems $T \Rightarrow S_i$, this gives $S_1 \wedge \cdots \wedge S_n \Leftrightarrow T$.

The completeness of a shadow set is verified in Lean by compiling checker theorems in both directions. For each shadow S_i, we compile a forward checker theorem establishing $T \Rightarrow S_i$. A candidate shadow is accepted only if its forward checker compiles. We then compile the backward checker theorem $S_1 \wedge \cdots \wedge S_n \Rightarrow T$. The backward checker theorem compiles only when the shadow statements jointly imply T, which is precisely the completeness condition illustrated in Figure 2.

The annotation process of complete shadow sets and corresponding checker theorems can be mostly automated (§4.1).

The following example shows common ways to build complete shadow sets. Additional checker patterns appear in Appendix K.

Example 1 (Bundled conclusions)**.** Suppose $A, B_1, \ldots, B_n$ are predicates and the intended formal statement has the form

$$A \Rightarrow B_1 \wedge \cdots \wedge B_n.$$

For each i, define a shadow theorem

$$S_i : A \Rightarrow B_i.$$

Then $\mathcal{S}$ is complete, because the individual components can be assembled back into the bundled conclusion.

Appendices K.1 and K.2 show the corresponding Lean checkers.

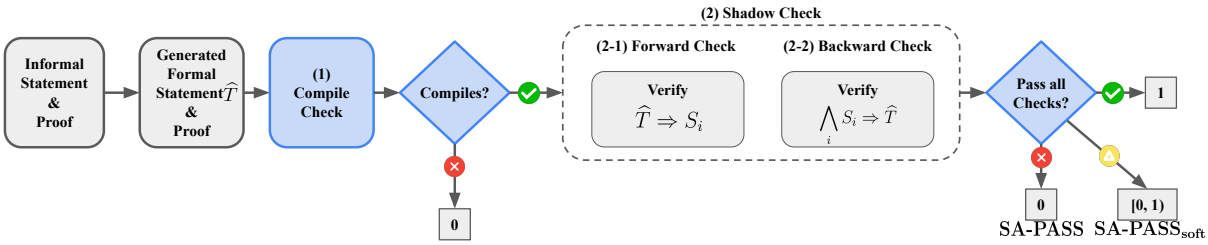


Figure 3: Evaluation procedure for SA-PASS and SA-PASS$_{\text{soft}}$ on one SHADOWBENCH problem.

3.3 SA-PASS Scoring

Definition 4 (Passing condition). *For a problem whose intended formal statement is T, a generated formal theorem with statement $\widehat{T}$ passes if and only if its declaration and proof type-check in Lean and it passes the forward checks*

$$\widehat{T} \Rightarrow S_1, \quad \widehat{T} \Rightarrow S_2, \quad \ldots, \quad \widehat{T} \Rightarrow S_n$$

and the backward check

$$S_1 \wedge \cdots \wedge S_n \Rightarrow \widehat{T},$$

where $\mathcal{S} = \{S_1, \ldots, S_n\}$ is a complete shadow set for T.

During evaluation, we append the forward and backward checkers to the generated Lean code, and compile the resulting file.

Because $\mathcal{S}$ is complete, passing all forward checks establishes $\widehat{T} \Rightarrow S_1 \wedge \cdots \wedge S_n \Rightarrow T$. Passing the backward check certifies $T \Rightarrow S_1 \wedge \cdots \wedge S_n \Rightarrow \widehat{T}$, and it rejects generated formal statements that add constraints not required by the intended formal statement. For example, if the generated formal statement has the form $\widehat{T} = T \wedge R$ for an arbitrary extra requirement R, then $\widehat{T} \Rightarrow T$ but $T \not\Rightarrow \widehat{T}$. Together, the forward and backward checks establish $\widehat{T} \Leftrightarrow T$.

The passing condition in Definition 4 induces the following binary semantic-alignment metric.

Definition 5 (SA-PASS). *For a generated formal theorem with statement $\widehat{T}$ and a complete shadow set $\mathcal{S} = \{S_1, \ldots, S_n\}$, define*

$$\text{SA-PASS}(\widehat{T}, \mathcal{S}) = \begin{cases} 1, & \begin{array}{l} \textit{if } \widehat{T} \textit{ compiles,} \\ \widehat{T} \Rightarrow S_i \textit{ for all } i, \\ \textit{and } \bigwedge_{i=1}^{n} S_i \Rightarrow \widehat{T}; \end{array} \\ 0, & \textit{otherwise.} \end{cases}$$

Since $\mathcal{S}$ is complete for the intended formal statement T, SA-PASS$(\widehat{T}, \mathcal{S})$=1 certifies $\widehat{T} \Leftrightarrow T$.

Definition 6 (SA-PASS$_{\text{soft}}$). *Let $\widehat{T}$ be the statement of a generated formal theorem, and let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a complete shadow set. If the generated theorem does not type-check in Lean, set* SA-PASS$_{\text{soft}}(\widehat{T}, \mathcal{S})$=0. *Otherwise, its* SA-PASS$_{\text{soft}}$ *score is*

$$\text{SA-PASS}_{\text{soft}}(\widehat{T}, \mathcal{S}) := \frac{1}{2} \cdot \underbrace{\frac{|\{i : \widehat{T} \Rightarrow S_i\}|}{|\mathcal{S}|}}_{\textit{forward checks}} + \frac{1}{2} \cdot \underbrace{\mathbf{1}\{\bigwedge_{i=1}^{n} S_i \Rightarrow \widehat{T}\}}_{\textit{backward check}}.$$

Here $\mathbf{1}\{\phi\}$ equals 1 if the corresponding checker theorem type-checks in Lean, and 0 otherwise.

A generated statement may become weaker than the intended theorem by adding assumptions, or stronger by omitting them. SA-PASS$_{\text{soft}}$ assigns partial credit in such cases by averaging the fraction of successful forward checks and the backward check indicator. By construction, SA-PASS$_{\text{soft}}(\widehat{T}, \mathcal{S}) = 1 \Leftrightarrow$ SA-PASS$(\widehat{T}, \mathcal{S}) = 1$.

Figure 3 summarizes the evaluation procedure.

4 SHADOWBENCH

To evaluate full autoformalization systems with SA-PASS, we construct SHADOWBENCH, a Lean 4 benchmark equipped with checker theorems for semantic alignment evaluation. Its checker construction is LLM-assisted and compiler-verified, while the resulting benchmark offers broad mathematical coverage and remains challenging for current systems.

LLM-assisted checker construction. Qwen3-235B proposes candidate shadow sets and their checker theorems, Lean verifies shadow set completeness, and experts review the results and guide revisions when needed. The checker construction takes about 10 minutes per problem on average, where for 85% of problems the initial LLM-generated shadow set is already complete (§4.1).

Area	L1	L2	L3	Total
`geometry` (Geo)	18	6	0	24
`topology` (Top)	16	9	0	25
`algebra` (Alg)	19	4	3	26
`analysis` (Anl)	14	10	4	28
`algebraic-geometry` (AG)	2	16	4	22
`combinatorics` (Cmb)	8	6	0	14
`number-theory` (NT)	17	0	2	19
`probability` (Prob)	19	1	0	20
Total	113	52	13	178

Table 1: Number of problems per area and difficulty level in SHADOWBENCH. One topology problem was withdrawn after an error was found in its reference statement.

Type	Count	Mean	Min	Max
Forward	513	2.88	1	13
Backward	193	1.08	1	12
Total	706	3.97	1	25

Table 2: Statistics of forward and backward checker theorems in SHADOWBENCH, counted over its 178 problems.

Coverage. SHADOWBENCH contains 178 problems ranging from postgraduate- to research-level across eight mathematical areas and three difficulty levels (Table 1). L1 and L2 contain postgraduate-level problems, separated by reference Lean length as a proxy for formalization complexity, with L2 containing longer formalizations. L3 contains research-level problems.

Difficulty. SHADOWBENCH remains challenging for current systems. Claude Code (Opus 4.8) with Numina-Lean-Agent, the strongest system we evaluate under SA-PASS, achieves a 61.8% compile rate but scores only 18.3% on SA-PASS$_{\text{soft}}$ and 11.2% on SA-PASS (Table 3).

4.1 Benchmark Construction

The curation interface for entering informal statements, reference Lean code, metadata, and checker blocks is shown in Appendix (Figures 8–10). Five mathematicians, spanning graduate students, postdoctoral researchers, and faculty, performed semantic review and guided revisions throughout the construction process.

Informal theorem. Annotators first collect source problems from textbooks, lecture notes, and research repositories, summarized in Table 8. When only a Lean formalization is available, Qwen3-235B drafts the informal theorem. We exclude problems whose solutions are dominated by theorem retrieval, since such cases test library search more than full autoformalization.

Reference formalization. For each informal theorem, we construct a reference Lean formalization, with Qwen3-235B providing a draft when only the informal theorem is available. Annotators edit the code until the theorem type-checks in Lean, verify that its formal statement matches the informal theorem, and use LeanSearch[3] to correct minor errors such as hallucinated theorem names.

Checker theorems. Given the reference formalization, Qwen3-235B proposes a candidate shadow set and its checker theorems (Figure 15), and Lean verifies the set's completeness by compiling all forward and backward checkers (§3.2). The initial draft is complete for 85% of problems. For the remaining 15%, an expert guides the model to revise or add shadows until all checkers compile. For 77.1% of checker proofs, `exact` alone suffices because the goal follows directly from existing library facts or the statement structure. Including LLM drafting, checker construction averages about 10 minutes per problem. Table 2 summarizes the number of forward and backward checker theorems per problem.

While each checker is constructed using the reference formalization, valid solver outputs may use different binder styles, declaration names, or field notation. To reduce false negatives, we diversify the checkers during benchmark construction to cover common differences in binder style and field notation. At evaluation time, rule-based rewriting first addresses any remaining declaration-name mismatches. If rule-based rewriting fails, Qwen3-235B attempts to generate an adapter proof that establishes equivalence between the solver declaration and the corresponding reference declaration, allowing the checker to run in Lean.

4.2 Benchmark Release

We provide the informal theorems and their formalization rules through a hosted portal such as CodaBench,[4] while keeping the shadow and checker theorems hidden to prevent direct optimization for the checks. A version-pinned evaluator runs the hidden checkers and reports the SA-PASS and SA-PASS$_{\text{soft}}$ scores.

[3] https://leansearch.net/
[4] https://www.codabench.org/

5 Experiment

5.1 Experimental Setup

Evaluation metrics. The main results in Table 3 use three metrics. **Compile rate** (Azerbayev et al., 2023) is the fraction of outputs whose final `theorem` block type-checks in the pinned Lean 4 and Mathlib environment.[5] **SA-PASS** (Definition 5) is 1 if the generated theorem type-checks and all forward and backward checks pass, and 0 otherwise. **SA-PASS$_{\text{soft}}$** (Definition 6) averages the forward-check pass rate and the backward-check indicator.

To analyze agreement with expert judgments (Tables 4 and 6), we include three additional metrics. **BLEU** (Azerbayev et al., 2023; Papineni et al., 2002) measures lexical n-gram overlap between the generated theorem block and the reference Lean theorem. **BEq+** (Poiroux et al., 2025b) uses Lean tactics to prove both directions of implication between the generated and reference statements. **LLM-as-judge** (Guo et al., 2025b) uses Gemini Pro 3.1, Claude Opus 4.7, and GPT-5.4. We queried each judge three times and took the majority over all nine votes, where the total API cost was approximately $480.

Methods. The main experimental results in Table 3 cover four method groups. For **agentic methods**, we run Claude Code with Claude Opus 4.8 (Anthropic, 2026b), Claude Opus 4.6 (Anthropic, 2026a), or Qwen3-235B (Yang et al., 2025), and Codex with GPT-5.4 (OpenAI, 2026a). Each configuration is evaluated with and without Numina-Lean-Agent (Liu et al., 2026a), which provides search and compilation tools during proof construction (Table 10). For **closed-source LLMs**, we use Claude Haiku 4.5, Claude Sonnet 4.6, and Claude Opus 4.6 (Anthropic, 2025, 2026c,a), Gemini Flash 2.5 and Gemini Pro 3.1 (Google DeepMind, 2025, 2026), and GPT-5.4 nano, GPT-5.4 mini, and GPT-5.4 (OpenAI, 2026a,b). For **open-source LLMs**, we run GPT-OSS-20B and GPT-OSS-120B (Agarwal et al., 2025), DeepSeek V3.2 and DeepSeek R1 0528 (DeepSeek-AI, 2025; Guo et al., 2025a), qwen3-coder and Qwen3-235B (Yang et al., 2025), and Llama 3.1 8B and 70B (Grattafiori et al., 2024). For the **Lean-specialized methods**, we use two-stage formalizer-to-prover configurations: Kimina 7B→8B (Wang et al., 2025; Numina and Kimi Team, 2025), Goedel 8B→8B, and Goedel 8B→32B (Lin et al., 2025b). Further model and configuration details are provided in Appendix B.2.

Method Configurations. Agentic methods use the default configuration of each agent, with prompts shown in Appendix I. Closed-source and open-source LLMs use zero-shot prompting with `max_tokens`=8192, temperature 0.6, and `top_p`=0.95. Lean-specialized methods use `max_tokens`=2048 for the formalizer and `max_tokens`=16384 for the prover, both with temperature 0.6 and `top_p`=0.95.

5.2 Results on SHADOWBENCH

Table 3 reports compile rate, SA-PASS$_{\text{soft}}$, and SA-PASS across all evaluated methods and difficulty levels.

Compilation substantially overestimates semantic alignment on SHADOWBENCH. The strongest system under SA-PASS, Claude Code (Opus 4.8) with Numina-Lean-Agent, reaches a 61.8% compile rate but only 18.3% SA-PASS$_{\text{soft}}$ and 11.2% SA-PASS. Thus, even the relaxed score remains 43.5 points below compile rate, while the gap to full semantic alignment is 50.6 points. This discrepancy primarily reflects false positives from compile rate, which achieves a precision of only 0.178 in our expert analysis (§6.1), compared with 1.000 for SA-PASS.

Numina-Lean-Agent improves compilability more than semantic alignment. Employing a concurrent agentic method Numina-Lean-Agent raises the compile rate of Claude Code (Opus 4.8) from 20.2% to 61.8%, a gain of 41.6 points. Over the same comparison, SA-PASS$_{\text{soft}}$ increases from 5.1% to 18.3% and SA-PASS from 2.8% to 11.2%, gains of 13.2 and 8.4 points, respectively. For Codex, the corresponding gains are 38.7 points in compile rate, 13.5 points in SA-PASS$_{\text{soft}}$, and 7.9 points in SA-PASS. The same qualitative pattern holds for the other agentic configurations. These results indicate that search and compiler feedback are substantially more effective at helping agents produce type-correct Lean theorem–proof pairs than at ensuring that the generated statements express the intended mathematics.

Semantic alignment on SHADOWBENCH remains largely unsolved. Closed-source LLMs, open-source LLMs, and Lean-specialized methods all score 0.0% on SA-PASS, and none exceeds

[5] v4.29.0

Method	L1 (n=113)			L2 (n=52)			L3 (n=13)			Average (n=178)		
	Compile	SA-PASS$_{soft}$	SA-PASS	Compile	SA-PASS$_{soft}$	SA-PASS	Compile	SA-PASS$_{soft}$	SA-PASS	Compile	SA-PASS$_{soft}$	SA-PASS
Agentic methods												
Claude Code (Opus 4.6)	6.2	1.8	0.0	1.9	1.0	0.0	0.0	0.0	0.0	4.5	1.4	0.0
+ Numina	55.8	**16.8**	**8.8**	36.5	17.3	11.5	15.4	0.0	0.0	47.2	15.7	9.0
Claude Code (Opus 4.8)	24.8	4.9	2.7	13.5	6.7	3.8	7.7	0.0	0.0	20.2	5.1	2.8
+ Numina	61.1	13.3	6.2	**71.2**	**31.7**	**23.1**	**30.8**	**7.7**	**7.7**	**61.8**	**18.3**	**11.2**
Claude Code (Qwen3 235B)	4.4	1.3	0.9	5.8	0.0	0.0	7.7	0.0	0.0	5.1	0.8	0.6
+ Numina	8.8	1.3	0.9	3.8	0.0	0.0	0.0	0.0	0.0	6.7	0.8	0.6
Codex (GPT-5.4)	20.4	4.4	2.7	11.5	1.9	1.9	7.7	0.0	0.0	16.9	3.4	2.2
+ Numina	**61.9**	15.9	**8.8**	48.1	21.2	13.5	**30.8**	**7.7**	**7.7**	55.6	16.9	10.1
Closed-source LLMs												
Claude Opus 4.6	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.0
Claude Sonnet 4.6	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0
Claude Haiku 4.5	0.9	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.3	0.0
GPT-5.4	1.8	0.4	0.0	5.8	0.0	0.0	7.7	0.0	0.0	3.4	0.3	0.0
GPT-5.4 mini	8.8	0.4	0.0	3.8	0.0	0.0	**30.8**	0.0	0.0	8.9	0.3	0.0
GPT-5.4 nano	6.1	0.0	0.0	5.8	0.0	0.0	15.4	0.0	0.0	6.7	0.0	0.0
Gemini 3.1 Pro	0.9	0.0	0.0	0.0	0.0	0.0	15.4	0.0	0.0	1.7	0.0	0.0
Gemini 2.5 Flash	0.9	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.3	0.0
Open-source LLMs												
GPT-OSS-120B	0.9	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.3	0.0
GPT-OSS-20B	4.4	0.4	0.0	1.9	0.0	0.0	0.0	0.0	0.0	3.4	0.3	0.0
DeepSeek R1 0528	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0
DeepSeek V3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3 235B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-coder	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama3.1 70b	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama3.1 8b	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lean-specialized methods												
Kimina 7B→8B	0.0	0.0	0.0	1.9	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
Goedel 8B→32B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Goedel 8B→8B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3: Per-difficulty pass rates (%) on SHADOWBENCH. *Compile* is the standalone Lean compile rate. SA-PASS$_{soft}$ measures the fraction of forward and backward checkers that pass. SA-PASS requires all forward and backward checkers to pass.

Metric	Precision	Recall	F1	Agreement
Compile	0.178	**1.000**	0.302	0.178
BLEU	0.234	0.674	0.347	0.550
LLM-as-judge	0.235	0.093	0.133	0.785
BEq+	0.400	0.186	0.254	<u>0.806</u>
SA-PASS$_{soft}$	<u>0.414</u>	<u>0.953</u>	<u>0.577</u>	0.752
SA-PASS	**1.000**	0.930	**0.964**	**0.988**

Table 4: Binary agreement of automatic metrics with expert judgment, averaged measurement over all six agentic configurations on SHADOWBENCH. Per-configuration results are in Appendix D. The best value per column is bold, and the second best is underlined.

0.3% on SA-PASS$_{soft}$. Even among agentic methods, the best scores are only 11.2% on SA-PASS and 18.3% on SA-PASS$_{soft}$. Performance is particularly limited on L3, where the best SA-PASS scores 7.7%. For comparison, §6.2 finds that compile rate and SA-PASS differ by only 2.1 points on average and at most 7.8 points across the 16 non-agentic models evaluated on ProofNet. This contrast suggests that the longer and more compositional targets in SHADOWBENCH expose semantic failures that are less visible on shorter, mostly single-conclusion formalizations.

6 Analysis

6.1 Agreement with Expert Judgment

The large gap between compile rate and SA-PASS in Table 3 admits two possible explanations: compilation may accept semantically misaligned formalizations, or SA-PASS may reject valid alternative formulations. We distinguish these explanations by comparing each automatic metric with expert judgments.

We collect expert judgments for compile-passed generations on SHADOWBENCH, where compilation and semantic alignment can disagree. Two mathematics experts independently assess whether each generated theorem is semantically aligned with the source problem and resolve disagreements through discussion. We analyze six configurations with sufficient compile-passed outputs: Claude Code (Opus 4.6), Claude Code (Qwen3-235B), and Codex (GPT-5.4), each with and without Numina-Lean-Agent. Against the resulting binary judgments, we measure precision, recall, F1, and agree-

Model	Compile	SA-PASS$_{soft}$	SA-PASS
Closed-source LLMs			
Claude Opus 4.6	**96.5**	92.9	90.6
Claude Sonnet 4.6	85.9	85.1	85.1
Claude Haiku 4.5	**96.5**	**95.3**	**95.3**
GPT 5.4	94.9	93.7	93.7
GPT 5.4 Mini	95.3	94.1	94.1
GPT 5.4 Nano	77.6	76.5	76.5
Gemini 3.1 Pro	35.7	34.9	34.9
Gemini 2.5 Flash	82.4	76.1	76.1
Open-source LLMs			
GPT-OSS-120B	60.8	60.8	60.8
GPT-OSS-20B	60.4	60.4	60.4
DeepSeek R1 0528	25.5	25.5	25.5
DeepSeek V3.2	64.3	64.3	64.3
Qwen3 235B	91.0	91.0	91.0
Qwen3-coder	89.8	89.8	89.8
Llama 3.1 70B	32.9	25.1	25.1
Llama 3.1 8B	35.3	28.2	28.2

Table 5: Non-agentic pass rates (%) on the Lean 4 version of ProofNet, using automatically generated shadow sets.

ment for compile rate, BLEU, BEq+, LLM-as-judge, SA-PASS$_{soft}$, and SA-PASS. For the non-binary metrics, we use thresholds of 0 for SA-PASS$_{soft}$ and 0.1 for BLEU, as determined in Appendix E. Agreement measures the proportion of an automatic metric's binary decisions that match expert judgments.

Table 4 shows that SA-PASS achieves the highest F1 (0.964) and agreement (0.988). Compile rate treats compilation success as semantically aligned, yielding 1.000 recall but only 0.178 precision due to false positives. BEq+ applies Lean's built-in equivalence check, achieving higher precision than Compile rate (0.400 vs. 0.178) but substantially lower recall (0.186). Its precision remains limited because the check covers only the final theorem statement, ignoring the auxiliary declarations required for a complete formalization of the informal theorem. In contrast, SA-PASS achieves 1.000 precision and 0.930 recall. Its few false negatives stem from namespace mismatches that the LLM fallback fails to resolve (Appendix D).

Taken together, these results indicate that the large gap between Compile rate and SA-PASS in Table 3 primarily reflects false positives from compilation rather than false negatives from SA-PASS.

6.2 SA-PASS on ProofNet

To test whether SA-PASS transfers beyond SHADOWBENCH, we apply it to the Lean 4 version of ProofNet (Azerbayev et al., 2023), an

Metric	Precision	Recall	F1	Agreement
Compile	0.800	**1.000**	<u>0.889</u>	<u>0.800</u>
BLEU	0.800	**1.000**	<u>0.889</u>	<u>0.800</u>
LLM-as-judge	0.865	0.889	0.877	<u>0.800</u>
BEq+	<u>0.887</u>	0.653	0.752	0.656
SA-PASS$_{soft}$	**1.000**	**1.000**	**1.000**	**1.000**
SA-PASS	**1.000**	**1.000**	**1.000**	**1.000**

Table 6: Binary agreement of automatic metrics with expert judgment for Llama 3.1 8B outputs on ProofNet. The best value per column is bold, and the second best is underlined.

undergraduate-level benchmark with shorter formalizations.[6] We construct a complete shadow set and checker theorems for each problem using the same LLM-assisted pipeline as for SHADOWBENCH (§4.1). All initial drafts pass the Lean completeness checks, so no expert revision is required.

We first compare the resulting metrics with expert judgments on Llama 3.1 8B outputs, which include both positive and negative expert labels. As shown in Table 6, SA-PASS and SA-PASS$_{soft}$ both match every expert judgment, while compile rate reaches an F1 of 0.889. This result provides evidence that the shadow checks remain reliable when applied to ProofNet.

Table 5 shows that compile rate and SA-PASS are much closer on ProofNet than on SHADOWBENCH. Across the 16 models, the gap is 2.1 points on average and at most 7.8 points. The smaller gap is consistent with the shorter and less compositional targets in ProofNet. Reference statements in SHADOWBENCH are about 1.6× longer by line count, and its reference solutions contain 4.5 auxiliary declarations per problem on average, compared with none in ProofNet (Table 9). Most ProofNet targets also have a single conclusion, so their shadow sets typically require only one or two short checks. Compile rate is therefore a closer proxy for semantic alignment on ProofNet, whereas SHADOWBENCH exposes failures that are less visible in shorter formalizations.

6.3 Forward and Backward Checker Analysis

The forward and backward checks used by SA-PASS detect different forms of semantic mismatch. A successful forward check certifies that the generated statement implies a required consequence of the intended theorem, while a successful back-

[6] We exclude the 31% of problems with known faulty reference formalizations (Poiroux et al., 2025b).

Method	L1 (n=113)				L2 (n=52)				L3 (n=13)				Average (n=178)			
	Fwd	Fwd_{all}	Bwd	SA-PASS	Fwd	Fwd_{all}	Bwd	SA-PASS	Fwd	Fwd_{all}	Bwd	SA-PASS	Fwd	Fwd_{all}	Bwd	SA-PASS
Claude Code (Opus 4.6)	0.9	0.9	2.7	0.0	1.9	1.9	0.0	0.0	0.0	0.0	0.0	0.0	1.1	1.1	1.7	0.0
+ Numina	15.0	15.0	**18.6**	**8.8**	17.3	17.3	17.3	11.5	0.0	0.0	0.0	0.0	14.6	14.6	16.9	9.0
Claude Code (Opus 4.8)	4.4	4.4	5.3	2.7	5.8	5.8	7.7	3.8	0.0	0.0	0.0	0.0	4.5	4.5	5.6	2.8
+ Numina	9.7	9.7	16.8	6.2	**25.0**	**25.0**	**38.5**	**23.1**	**7.7**	**7.7**	**7.7**	**7.7**	14.0	14.0	**22.5**	**11.2**
Claude Code (Qwen3 235B)	0.9	0.9	1.8	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.6	1.1	0.6
+ Numina	0.9	0.9	1.8	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.6	1.1	0.6
Codex (GPT-5.4)	3.5	3.5	5.3	2.7	1.9	1.9	1.9	1.9	0.0	0.0	0.0	0.0	2.8	2.8	3.9	2.2
+ Numina	**17.7**	**17.7**	14.2	**8.8**	19.2	19.2	23.1	13.5	**7.7**	**7.7**	**7.7**	**7.7**	**17.4**	**17.4**	16.3	10.1

Table 7: Per-difficulty checker metrics (%) for agentic systems on SHADOWBENCH. Fwd and Bwd average the per-problem fractions of passed forward and backward checkers, Fwd_{all} is the fraction of problems passing every forward checker.

ward check certifies that the intended theorem implies the generated statement. Passing the backward checks but not all forward checks can arise from omitted conclusions or added assumptions (Appendix J.4). Conversely, passing all forward checks but failing the backward checks can arise from added conclusions or omitted assumptions. When both directions fail, neither implication is certified (Appendix J.5).

Table 7 shows that the two implication directions can differ substantially even when SA-PASS is low. Claude Code (Opus 4.8) with Numina-Lean-Agent reaches a SA-PASS score of 11.2%, with a 14.0% all forward pass rate but 22.5% average backward pass rate. Among its 110 compile-passed outputs, 20 pass both directions, 20 pass only the backward direction, 5 pass only the forward direction, and 65 pass neither.

7 Conclusion

We introduced SA-PASS, an automatic metric that evaluates semantic alignment between generated formal statements and informal theorem statements. For each problem, SA-PASS uses a set of auxiliary formal statements, called shadows, whose conjunction is equivalent to the intended theorem statement, and checks in Lean whether the generated statement implies each shadow and is implied by their conjunction. We instantiated SA-PASS in SHADOWBENCH, a Lean 4 full autoformalization benchmark of 178 postgraduate- to research-level problems spanning eight mathematical areas. On SHADOWBENCH, Claude Code (Opus 4.8) with Numina-Lean-Agent reaches 61.8% compile rate, 18.3% $\text{SA-PASS}_{\text{soft}}$, and 11.2% SA-PASS, showing that compile success alone can overstate semantic alignment. Across six agentic configurations, SA-PASS achieves 98.8% binary agreement with expert judgments.

Limitations

SHADOWBENCH covers 178 postgraduate- to research-level problems across eight mathematical areas, and future versions can extend this coverage to more domains and proof styles. SA-PASS requires one-time construction of shadow statements and checker proofs for each benchmark problem.

Potential Risks

The main risk is benchmark leakage. If checker artifacts are exposed, systems can optimize for the checks rather than the informal theorem. To reduce this risk, public problem inputs are separated from the checking artifacts used for evaluation. We give the full release and evaluation protocol in §4.2.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-00520280), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], Electronics and Telecommunications Research Institute (ETRI) grant funded by ICT R&D program of MSIT/IITP (2022-0-00995, Automated reliable source code generation from natural language descriptions). We thank Soonho Kong for supporting our experiments with Claude and Claude Code on Amazon Web Services.

References

1000+ Theorems contributors. 2026. 1000+ Theorems: Euler's quadrilateral theorem. https://1000-plus.github.io/all. Entry "Euler's quadrilateral theorem"; accessed 2026-05-26.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Manindra Agrawal, Neeraj Kayal, and Nitin Saxena. 2004. Primes is in p. *Annals of mathematics*, pages 781–793.

Martin Aigner and Günter M Ziegler. 1999. Proofs from the book. *Berlin. Germany*, 1(2):7.

Pawan Sasanka Ammanamanchi, Siddharth Bhat, and Stella Biderman. 2026. Faults in our formal benchmarking: Dataset defects and evaluation failures in Lean theorem proving. In *Forty-third International Conference on Machine Learning*.

Anthropic. 2025. System Card: Claude Haiku 4.5.

Anthropic. 2026a. System Card: Claude Opus 4.6. *Claude Opus 4.6 Model Card*.

Anthropic. 2026b. System Card: Claude Opus 4.8. *Claude Opus 4.8 Model Card*.

Anthropic. 2026c. System Card: Claude Sonnet 4.6. *Claude Sonnet 4.6 Model Card*.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Guoxin Chen, Jing Wu, Xinjie Chen, Wayne Xin Zhao, Ruihua Song, Chengxi Li, Kai Fan, Dayiheng Liu, and Minpeng Liao. 2026. ReForm: Reflective autoformalization with prospective bounded sequence optimization. In *International Conference on Learning Representations*. Poster.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

D.A. Cox, J. Little, and D. O'Shea. 2008. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Undergraduate Texts in Mathematics. Springer New York.

DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Jean Alexandre Dieudonne and Alexandre Grothendieck. 1971. *Éléments de géométrie algébrique*, volume 166. Springer Berlin Heidelberg New York.

Yue Fang, Shaohan Huang, Xin Yu, Haizhen Huang, Zihan Zhang, Weiwei Deng, Furu Wei, Feng Sun, Qi Zhang, and Zhi Jin. 2025. NL2Lean: Translating natural language into Lean 4 through multi-aspect reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31148–31158, Suzhou, China. Association for Computational Linguistics.

Guoxiong Gao, Yutong Wang, Jiedong Jiang, Qi Gao, Zihan Qin, Tianyi Xu, and Bin Dong. 2025. Herald: A natural language annotated Lean 4 dataset. In *International Conference on Learning Representations*. Poster.

Google DeepMind. 2025. Gemini 2.5 Flash Model Card.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. *Gemini 3.1 Pro Model Card*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025a. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638.

Qi Guo, Jianing Wang, Jianfei Zhang, Deyang Kong, Xiangzhou Huang, Xiangyu Xi, Wei Wang, Jingang Wang, Xunliang Cai, Shikun Zhang, and Wei Ye. 2025b. Autoformalizer with tool feedback. *arXiv preprint arXiv:2510.06857*.

Prithwish Jana, Kaan Kale, Ahmet Ege Tanriverdi, Cruise Song, Sriram Vishwanath, and Vijay Ganesh. 2026. ProofBridge: Auto-formalization of natural language proofs in Lean via joint embeddings. In *International Conference on Learning Representations*. Poster.

Jongyoon Kim, Hojae Han, and Seung-Won Hwang. 2026. Benchmarking testing in automated theorem proving. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 2241–2260, San Diego, California, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.

Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

John M Lee. 2003. Smooth manifolds. In *Introduction to smooth manifolds*, pages 1–29. Springer.

Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, and Chi Jin. 2025a. Goedel-Prover: A frontier model for open-source automated theorem proving. In *Conference on Language Modeling*.

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. 2025b. Goedel-Prover-V2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv preprint arXiv:2508.03613*.

Junqi Liu, Zihao Zhou, Zekai Zhu, Marco Dos Santos, Weikun He, Jiawei Liu, Ran Wang, Yunzhou Xie, Junqiao Zhao, Qiufeng Wang, Lihong Zhi, Jia Li, and Wenda Li. 2026a. Numina-Lean-Agent: An open and general agentic reasoning system for formal mathematics. *arXiv preprint arXiv:2601.14027*.

Qi Liu, Xinhao Zheng, Xudong Lu, Qinxiang Cao, and Junchi Yan. 2025a. Rethinking and improving autoformalization: Towards a faithful metric and a dependency retrieval-based approach. In *International Conference on Learning Representations*. Spotlight.

Xiaoyang Liu, Kangjie Bao, Jiashuo Zhang, Yunqi Liu, Yu Chen, Yuntian Liu, Yang Jiao, and Tao Luo. 2025b. ATLAS: Autoformalizing theorems through lifting, augmentation, and synthesis of data. In *Advances in Neural Information Processing Systems*.

Xiaoyang Liu, Tao Zhu, Zineng Dong, Yuntian Liu, Qingfeng Guo, Zhaoxuan Liu, Yu Chen, and Tao Luo. 2026b. ASSESS: A semantic and structural evaluation framework for statement similarity. In *International Conference on Learning Representations*.

Jianqiao Lu, Yingjia Wan, Yinya Huang, Jing Xiong, Zhengying Liu, and Zhijiang Guo. 2025a. FormalAlign: Automated alignment evaluation for autoformalization. In *International Conference on Learning Representations*.

Jianqiao Lu, Yingjia Wan, Zhengying Liu, Yinya Huang, Jing Xiong, Chengwu Liu, Jianhao Shen, Hui Jin, Jipeng Zhang, Haiming Wang, Zhicheng Yang, Jing Tang, and Zhijiang Guo. 2024. Process-driven autoformalization in Lean 4. *arXiv preprint arXiv:2406.01940*.

Wangyue Lu, Lun Du, Sirui Li, Ke Weng, Haozhe Sun, Hengyu Liu, Minghe Yu, Tiancheng Zhang, and Ge Yu. 2025b. Automated formalization via conceptual retrieval-augmented LLMs. *arXiv preprint arXiv:2508.06931*.

Logan Murphy, Kaiyu Yang, Jialiang Sun, Zhaoyu Li, Anima Anandkumar, and Xujie Si. 2024. Autoformalizing Euclidean geometry. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 36847–36893. PMLR.

Numina and Kimi Team. 2025. Kimina-Prover: Applying test-time RL search on large formal reasoning models.

OpenAI. 2026a. GPT-5.4 Thinking System Card.

OpenAI. 2026b. Introducing GPT-5.4 mini and nano.

Azim Ospanov, Farzan Farnia, and Roozbeh Yousefzadeh. 2025. miniF2F-Lean revisited: Reviewing limitations and charting a path forward. In *Advances in Neural Information Processing Systems*. Poster.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Auguste Poiroux, Antoine Bosselut, and Viktor Kunčak. 2025a. RLMEval: Evaluating research-level neural theorem proving. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10946–10957, Suzhou, China. Association for Computational Linguistics.

Auguste Poiroux, Gail Weiss, Viktor Kunčak, and Antoine Bosselut. 2025b. Reliable evaluation and benchmarks for statement autoformalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17947–17969, Suzhou, China. Association for Computational Linguistics.

Jacob Shapiro. 2024. Functional analysis: Princeton university mat520 lecture notes. https://web.math.princeton.edu/~js129/PDFs/teaching/MAT520_fall_2023/MAT520_Lecture_Notes.pdf. Created August 18, 2023; last typeset September 5, 2024; accessed May 26, 2026.

The Stacks project authors. 2026. The stacks project. https://stacks.math.columbia.edu.

Elias M Stein and Rami Shakarchi. 2009. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press.

The mathlib Community. 2020. The Lean Mathematical Library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, CPP 2020, New Orleans, LA, USA. ACM.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, Jianqiao Lu, Hugues de Saxcé, Bolton Bailey, Chendong Song, Chenjun Xiao, Dehao Zhang, Ebony Zhang, Frederick Pu, Han Zhu, and 21 others. 2025. Kimina-Prover preview: Towards large formal reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.11354*.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. In *Advances in Neural Information Processing Systems*.

A David Wunsch. 2005. *Complex variables with applications*. Pearson Education India.

Jiaxuan Xie, Chengwu Liu, Ye Yuan, Siqi Li, Zhiping Xiao, and Ming Zhang. 2025. FMC: Formalization of natural language mathematical competition problems. In *2nd AI for Math Workshop at ICML 2025*.

Huajian Xin, Daya Guo, Zhihong Shao, Z. Z. Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. DeepSeek-Prover: Advancing theorem proving in LLMs through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems*.

Huaiyuan Ying, Zijian Wu, Yihan Geng, Jiayu Wang, Dahua Lin, and Kai Chen. 2024. Lean Workbook: A large-scale Lean problem set formalized from natural language math problems. In *Advances in Neural Information Processing Systems*. Datasets and Benchmarks Track Poster.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. miniF2F: A cross-system benchmark for formal Olympiad-level mathematics. In *International Conference on Learning Representations*.

A Dataset Details

Dataset		SHADOWBENCH	ProofNet
n		178	371
Target statement	Lines	5.1	3.2
	Chars	226	149
Reference proof	Lines	71.6	3.4
	Chars	3242	157
	Aux. decl.	4.5	0.0

Table 9: Statistics of the reference statement and proof for both SHADOWBENCH and ProofNet (Lean 4).

A.1 Problem Sources

Table 8 lists the sources used to construct SHADOWBENCH. Annotators collect problems from textbooks, lecture notes, research papers, and Lean 4 library material. The Informal and Formal columns mark the statement type available in each source before annotation. When a source provides only one side, Qwen3-235B drafts the other, as described in §4.1.

A.2 Dataset Statistics

Table 9 compares reference Lean statistics of SHADOWBENCH with the Lean 4 version of ProofNet. Target statements in SHADOWBENCH are about $1.6\times$ longer by line count than those in ProofNet. The larger difference lies in the reference proofs. A reference proof in SHADOWBENCH averages 72 lines and 4.5 auxiliary declarations, while ProofNet averages about 3.4 lines and none of auxiliary declarations. Table 2 reports the number of checker theorems per problem.

B Experiment Details

B.1 Generation Setup

Closed-source models and general-purpose open-source models are accessed through OpenRouter. Lean-specialized models are self-hosted with vLLM (Kwon et al., 2023). These models are executed on two NVIDIA Titan RTX (24GB) for model sizes up to 8B, on two NVIDIA RTX 6000 PRO Ada Generation (48GB) for the 32B model, and on eight NVIDIA RTX 6000 PRO Black Edition (96GB) for the 235B model. All non-agentic runs use zero-shot prompting with temperature 0.6. For each non-agentic problem run, we generate a single completion. For the agentic setting, each harness is evaluated with and without Numina-Lean-Agent (Liu et al., 2026a).

B.2 Model Details

Base system	Backbone	Thinking
Claude Code	Claude Opus 4.6 (Anthropic, 2026a)	Medium
Claude Code	Claude Opus 4.8 (Anthropic, 2026b)	Medium
Claude Code	Qwen3 235B (Yang et al., 2025)	-
Codex	GPT-5.4 (OpenAI, 2026a)	Medium

Table 10: Configuration for agentic harnesses.

We evaluate 19 non-agentic model configurations across 9 model families, as listed in Table 11. Parameter counts for closed-source models are not publicly available. General-purpose open-source models are evaluated directly, while Lean-specialized rows use the two-stage formalization-to-proving setup shown in Table 11. The four agentic settings are Claude Code with Claude Opus 4.6, or Claude Opus 4.8, Claude Code with Qwen3 235B, and Codex with GPT-5.4. Each setting is evaluated both with and without Numina-Lean-Agent, as listed in Table 10.

Source	Reference	Informal	Formal
Analysis problem	(Shapiro, 2024)	O	X
Complex Variables with Applications	(Wunsch, 2005)	O	X
Convex Optimization	(Boyd and Vandenberghe, 2004)	O	X
Elements de geometrie algebrique (EGA)	(Dieudonne and Grothendieck, 1971)	O	X
Ideals, Varieties, and Algorithms (4th ed.)	(Cox et al., 2008)	O	X
Introduction to Smooth Manifolds	(Lee, 2003)	O	X
Mathlib 4	(The mathlib Community, 2020)	X	O
Missing theorems from Wiedijk 1000+	(1000+ Theorems contributors, 2026)	O	X
PRIMES is in P	(Agrawal et al., 2004)	O	X
Proofs from THE BOOK	(Aigner and Ziegler, 1999)	O	X
Real Analysis: Measure Theory, Integration, and Hilbert Spaces	(Stein and Shakarchi, 2009)	O	X
The Stacks Project	(Stacks project authors, 2026)	△	△

Table 8: List of reference sources used to construct SHADOWBENCH. *Informal* and *Formal* mark the statement type available before annotation. A triangle denotes mixed availability across problems.

Setting	Family	Models
Closed-source	Claude (Anthropic, 2025, 2026c,a)	Claude Haiku 4.5, Claude Sonnet 4.6, Claude Opus 4.6
	Gemini (Google DeepMind, 2025, 2026)	Gemini Flash 2.5, Gemini Pro 3.1
	GPT (OpenAI, 2026a)	GPT-5.4 nano, GPT-5.4 mini, GPT-5.4
Open-source general	GPT-OSS (Agarwal et al., 2025)	GPT-OSS-20B, GPT-OSS-120B
	DeepSeek (Guo et al., 2025a)	DeepSeek V3.2, DeepSeek R1 0528
	Qwen (Yang et al., 2025)	qwen3-coder, Qwen3 235B
	Llama (Grattafiori et al., 2024)	Llama 3.1 8B, Llama 3.1 70B
Open-source Lean-specialized	Kimina (Wang et al., 2025)	Kimina 7B→8B
	Goedel (Lin et al., 2025b)	Goedel 8B→8B, Goedel 8B→32B

Table 11: Non-agentic model configurations evaluated in our experiments.

B.3 Per-Metric Rates

This section expands §6.1. Table 12 reports BEq+, BLEU, and LLM-as-judge by difficulty. The judge prompt is in Appendix I.1 and the cost profile in Appendix C.3.

C Cost and Compute

This section reports the compute and API budget used for the experiments. Measured costs come from per-call sidecar logs with `prompt_tokens`, `completion_tokens`, and OpenRouter-quoted `cost_usd`. Estimated costs use observed prompt and completion lengths with provider pricing. Self-hosted Lean-specialized models and Qwen3 235B were served on local GPUs, so we report wall-clock time rather than cost.

C.1 Non-Agentic Generation

Table 13 aggregates per-call sidecar logs across the 33 non-agentic runs. Closed-source rows are billed by the model provider. Open-source API rows are billed through OpenRouter. Open-source self-hosted rows are local vLLM deployments and have zero API cost.

C.2 Agentic Generation

The agentic runs use Claude Code and Codex, each with and without Numina-Lean-Agent. Early experiments used subscription access, while later runs incurred paid API charges. For consistent accounting, Table 14 reports the estimated cost if all API-capable agentic runs were billed through API. We estimate cost from per-task wall-clock time and provider token throughput. The table reports wall-clock time and estimated API cost for the 178-problem run. Claude Code with Numina-Lean-Agent is the most costly API-billed setting at 34.1 minutes per problem. Claude Code with Qwen3 235B uses the local vLLM server, so it has no API charge in this table.

C.3 LLM-as-Judge

We score every cached generation with a three-judge majority vote (3 samples per judge, temperature 0.5). The three judges are `anthropic/claude-opus-4.7`, `google/gemini-3.1-pro-preview`, and `openai/gpt-5.4`, accessed via OpenRouter. At three samples per judge, each candidate triggers nine completions.

C.4 Total

Adding the three components above, the API spend for the experiments reported in the main body is approximately \$3,970 (\$62 NA + \$3,430 agentic + \$480 judge). Self-hosted local GPU runs are not included in this API total.

D Per-Configuration Agreement with Expert Judgment

For expert judgment, two mathematics experts independently evaluate whether each generated theorem is semantically aligned with the source problem. The evaluation is finalized once both annotators reach consensus. The disagreements are resolved through discussion.

Table 4 in the main text pools all six agentic configurations. Tables 16–21 report them individually. Across every configuration, SA-PASS achieves a Precision of 1.0 against expert judgment, with Recall of 1.0 except on two of the three Numina configurations, where it is 0.889 and 0.947. Claude Code (Opus 4.6) without Numina-Lean-Agent produced only incorrect outputs, so Recall and F1 are undefined for it and Agreement is reported instead.

False Negatives of SA-PASS. SA-PASS misses three of the outputs the experts judged aligned.

Model	L1 (n=113)			L2 (n=52)			L3 (n=13)			Average (n=178)		
	BEq+	BLEU	Judge	BEq+	BLEU	Judge	BEq+	BLEU	Judge	BEq+	BLEU	Judge
Agentic methods												
Claude Code (Opus 4.6)	1.8	6.6	4.4	0.0	1.6	1.9	0.0	0.8	0.0	1.1	4.7	3.4
+ Numina	12.4	18.9	8.0	15.4	13.8	1.9	7.7	4.5	0.0	12.9	16.4	5.6
Claude Code (Qwen3 235B)	0.0	11.6	0.0	0.0	7.5	0.0	0.0	1.2	0.0	0.0	9.6	0.0
+ Numina	0.0	9.6	0.0	0.0	7.4	0.0	0.0	2.8	0.0	0.0	8.5	0.0
Codex (GPT-5.4)	1.8	12.4	5.3	0.0	10.2	1.9	7.7	3.8	0.0	1.7	11.1	3.9
+ Numina	9.7	15.6	5.3	7.7	11.8	1.9	7.7	2.7	0.0	9.0	13.5	3.9
Closed-source LLMs												
GPT-5.4	2.7	15.1	1.8	0.0	12.7	0.0	7.7	12.8	0.0	2.2	14.2	1.1
GPT-5.4 mini	2.7	11.8	0.0	0.0	11.5	1.9	7.7	12.5	0.0	2.2	11.8	0.6
GPT-5.4 nano	0.9	10.5	0.9	0.0	8.3	0.0	0.0	6.6	0.0	0.6	9.6	0.6
Gemini 3.1 Pro	4.4	13.8	5.3	3.8	11.7	0.0	7.7	14.4	0.0	4.5	13.2	3.4
Gemini 2.5 Flash	3.5	17.7	1.8	5.8	13.8	0.0	15.4	13.4	0.0	5.1	16.2	1.1
Claude Opus 4.6	6.2	19.8	1.8	1.9	18.0	1.9	7.7	14.7	0.0	5.1	18.9	1.7
Claude Sonnet 4.6	5.3	16.8	2.7	3.8	13.8	0.0	15.4	12.7	0.0	5.6	15.6	1.7
Claude Haiku 4.5	2.7	16.1	0.0	1.9	14.0	0.0	15.4	12.6	0.0	3.4	15.2	0.0
Open-source LLMs												
GPT-OSS-120B	1.8	13.9	2.7	1.9	13.1	0.0	7.7	14.2	0.0	2.2	13.7	1.7
GPT-OSS-20B	0.9	5.7	2.7	1.9	6.7	0.0	0.0	8.2	0.0	1.1	6.2	1.7
DeepSeek R1 0528	0.0	3.8	0.9	0.0	2.8	0.0	0.0	2.9	0.0	0.0	3.5	0.6
DeepSeek V3.2	3.5	16.2	0.0	0.0	11.9	0.0	15.4	11.6	0.0	3.4	14.6	0.0
Qwen3 235B	–	15.3	–	–	8.7	–	–	2.4	–	–	12.4	–
qwen3-coder	0.9	16.4	0.0	1.9	10.8	0.0	15.4	7.0	0.0	2.2	14.1	0.0
llama3.1 70b	0.0	10.2	0.0	1.9	8.4	0.0	7.7	6.6	0.0	1.1	9.4	0.0
llama3.1 8b	0.0	7.6	0.0	0.0	6.1	0.0	7.7	4.9	0.0	0.6	7.0	0.0
Lean-specialized methods												
Kimina 7B→8B	0.0	5.9	0.0	0.0	5.4	0.0	0.0	9.6	0.0	0.0	6.0	0.0
Goedel 8B→32B	0.0	6.6	0.0	0.0	6.0	0.0	0.0	1.7	0.0	0.0	6.1	0.0
Goedel 8B→8B	0.0	6.1	0.0	1.9	8.9	0.0	7.7	14.6	0.0	1.1	7.5	0.0

Table 12: Per-difficulty automatic metric results (%) on SHADOWBENCH. BEq+ and Judge are pass rates. BLEU is a token-level similarity score. The Average columns are task-weighted over 178 problems.

Bucket	max_tokens 8192	max_tokens 16384	Subtotal
Closed (API)	$23.10	$30.46	$53.56
Open (via OpenRouter)	$3.79	$4.70	$8.49
Open (self-hosted)	$0.00	$0.00	$0.00
Total	$26.89	$35.16	$62.05

Table 13: Non-agentic generation cost across 33 runs and 178 prompts per run, taken from per-call sidecar logs.

All three state the intended theorem correctly but name their auxiliary declarations differently from the reference. Two rename or omit them, and the other one declares the expected name as a theorem where the reference declares a definition. The LLM matching step of checker adaptation (§4.1) did not resolve these cases, so every forward check for those problems fails.

E Threshold Sensitivity

The relaxed score SA-PASS$_{\text{soft}}$ is the fraction of passing shadow checks, and is thus non-binary. To test whether the choice of threshold affects agreement with expert judgment, we sweep a threshold t and count a statement as positive when its passing fraction exceeds t (Figure 4). F1 is flat below $t = 0.5$, at 0.552 for Claude Code (Opus 4.6) with Numina-Lean-Agent and 0.623 for Codex with Numina-Lean-Agent, and rises to 0.941 and 0.973 at $t = 0.5$. A passing fraction above one half requires both directions to pass, so SA-PASS$_{\text{soft}}$

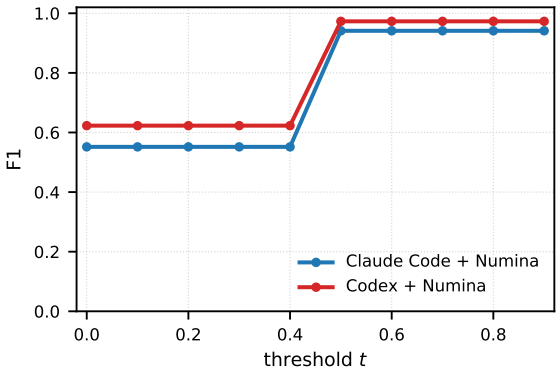


Figure 4: F1 of SA-PASS$_{\text{soft}}$ against expert judgment across the binarization threshold t.

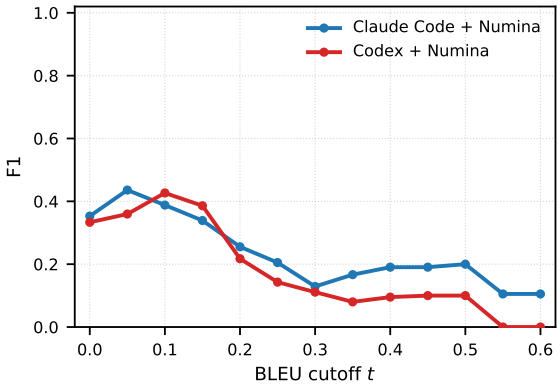


Figure 5: F1 of BLEU against expert judgment across the cutoff t.

coincides with SA-PASS from that point on. We therefore report SA-PASS$_{\text{soft}}$ with $t = 0$, the setting most distinct from SA-PASS.

For BLEU, F1 peaks at a low cutoff and then falls (Figure 5). Claude Code (Opus 4.6) with Numina-Lean-Agent peaks at 0.436 with a cutoff of 0.05, and Codex with Numina-Lean-Agent at 0.427 with a cutoff of 0.1. We report BLEU at a cutoff of 0.1, which maximizes the mean F1 over the two configurations that have expert-labelled positives, so BLEU is scored at its most favorable setting. No cutoff brings BLEU near the agreement SA-PASS reaches.

F Rank Correlation with Expert Judgment

We measure whether each automatic metric ranks whole systems in the same order as a human expert. Two human experts label every candidate Lean statement as `ok` or `reject`. We evaluate six agentic systems: Claude Code and Codex, each run directly and with Numina-Lean-Agent, and Qwen3 235B run directly and with Numina-Lean-Agent. For each system, we aggregate every score over all 178 tasks and rank the systems by the expert pass rate. Table 22 reports Spearman ρ and Kendall τ between the ranking of the six systems induced by each metric and the expert ranking.

SA-PASS recovers the expert ranking exactly ($\rho = \tau = 1.00$), and compile rate is next at $\rho = 0.99$, ahead of BLEU (0.90), SA-PASS$_{\text{soft}}$ (0.82), BEq+ (0.76), and LLM-as-judge (0.72). With six systems whose compile rates range from 4.5% to 55.6%, the system-level ranking is a coarse test, and the per-task agreement in Table 4 separates the metrics further.

G Public Challenge Track

We released an early 126-problem version of SHADOWBENCH as Track 4 of the ICML 2026 AI4Math Challenge. Figure 6 shows the CodaBench competition page that hosted the 126-problem snapshot of SHADOWBENCH as Track 4 of the ICML 2026 AI4Math Challenge. As we described in §4.2, informal theorems and formalization rules are accessible to the public, while the checker theorems for evaluation remain hidden. About 117 teams registered, but only twenty-seven teams actually participated.

Experiment	N tasks	Wall-clock (h)	Avg/task (min)	Est. cost (USD)
Claude Code (Opus 4.6)	178	20.5	6.9	$310
+ Numina	178	101.2	34.1	$1,500
Claude Code (Opus 4.8)	178	20.4	6.9	$310
+ Numina	178	51.8	17.5	$780
Codex (GPT-5.4)	178	2.8	0.9	$200
+ Numina	178	21.1	7.1	$330
Claude Code (Qwen3 235B)	178	–	–	local GPU
+ Numina	178	–	–	local GPU
Total		217.8		$3,430

Table 14: Estimated API cost of agentic methods over the 178 problems of SHADOWBENCH.

Figure 6: CodaBench page of SHADOWBENCH as Track 4 of the ICML 2026 AI4Math Challenge.

Model	Calls	Cost
anthropic/claude-opus-4.7	9,207	$324
google/gemini-3.1-pro-preview	9,207	$58
openai/gpt-5.4	9,207	$98
Total judge spend	27,621	$480

Table 15: Cost of LLM-as-judge.

Metric	Precision	Recall	F1	Agreement
Compile	0.214	**1.000**	0.353	0.214
BLEU	0.265	0.722	0.388	0.512
LLM-as-judge	0.375	0.167	0.231	<u>0.762</u>
BEq+	0.364	0.222	0.276	0.750
SA-PASS$_{\text{soft}}$	<u>0.400</u>	<u>0.889</u>	<u>0.552</u>	0.690
SA-PASS	**1.000**	<u>0.889</u>	**0.941**	**0.976**

Table 16: Binary agreement of automatic metrics with expert judgment for outputs from Claude Code (Opus 4.6) with Numina-Lean-Agent on SHADOWBENCH. The best value per column is bold, and the second best is underlined.

The participants' results show similar trends to our experiments. The SA-PASS$_{\text{soft}}$ distribution is skewed, where the median is 8.1%, and the average is 11.1%, while the compile rate averages 61.6%. That is, there is about a $5.5\times$ gap between the compile rate and SA-PASS$_{\text{soft}}$, underlining that the compile rate causes false positives in formalizations. In many of these low-scoring cases, the submission replaces the target with a trivial statement closed by `trivial`, which compiles but cannot be used by another theorem or declaration.

Metric	Precision	Recall	F1	Agreement
Compile	0.000	–	–	0.000
BLEU	0.000	–	–	0.250
LLM-as-judge	0.000	–	–	0.750
BEq+	–	–	–	**1.000**
SA-PASS$_{\text{soft}}$	0.000	–	–	0.375
SA-PASS	–	–	–	**1.000**

Table 17: Binary agreement of automatic metrics with expert judgment for outputs from Claude Code (Opus 4.6) on SHADOWBENCH. The best value per column is bold, and the second best is underlined.

Metric	Precision	Recall	F1	Agreement
Compile	0.192	**1.000**	0.322	0.192
BLEU	0.286	0.842	0.427	0.566
LLM-as-judge	0.143	0.053	0.077	0.758
BEq+	0.444	0.211	0.286	0.798
SA-PASS$_{\text{soft}}$	0.452	**1.000**	0.623	0.768
SA-PASS	**1.000**	0.947	**0.973**	**0.990**

Table 18: Binary agreement of automatic metrics with expert judgment for outputs from Codex (GPT-5.4) with Numina-Lean-Agent on SHADOWBENCH. The best value per column is bold, and the second best is underlined.

Metric	Precision	Recall	F1	Agreement
Compile	0.133	**1.000**	0.235	0.133
BLEU	0.000	0.000	0.000	0.567
LLM-as-judge	–	0.000	–	0.867
BEq+	–	0.000	–	0.867
SA-PASS$_{\text{soft}}$	0.500	**1.000**	0.667	0.867
SA-PASS	**1.000**	**1.000**	**1.000**	**1.000**

Table 19: Binary agreement of automatic metrics with expert judgment for outputs from Codex (GPT-5.4) on SHADOWBENCH. The best value per column is bold, and the second best is underlined.

Metric	Precision	Recall	F1	Agreement
Compile	0.083	**1.000**	0.154	0.083
BLEU	0.000	0.000	0.000	0.583
LLM-as-judge	–	0.000	–	0.917
BEq+	–	0.000	–	0.917
SA-PASS$_{\text{soft}}$	0.500	**1.000**	0.667	0.917
SA-PASS	**1.000**	**1.000**	**1.000**	**1.000**

Table 20: Binary agreement of automatic metrics with expert judgment for outputs from Claude Code (Qwen3 235B) with Numina-Lean-Agent on SHADOWBENCH. The best value per column is bold, and the second best is underlined.

Such submissions give the compile rate many false positives. They can even induce a ranking error, where the top five by compile rate share no member with the top five by SA-PASS$_{\text{soft}}$. Compilation and

Metric	Precision	Recall	F1	Agreement
Compile	0.111	**1.000**	0.200	0.111
BLEU	–	0.000	–	0.889
LLM-as-judge	–	0.000	–	0.889
BEq+	–	0.000	–	0.889
SA-PASS$_{\text{soft}}$	0.500	**1.000**	0.667	0.889
SA-PASS	**1.000**	**1.000**	**1.000**	**1.000**

Table 21: Binary agreement of automatic metrics with expert judgment for outputs from Claude Code (Qwen3 235B) on SHADOWBENCH. The best value per column is bold, and the second best is underlined.

Metric	ρ (↑)	τ (↑)
Compile	0.99	0.97
BLEU	0.90	0.79
LLM-as-judge	0.72	0.52
BEq+	0.76	0.57
SA-PASS$_{\text{soft}}$	0.82	0.71
SA-PASS	**1.00**	**1.00**

Table 22: Spearman ρ and Kendall τ between the ranking of the six agentic systems by each automatic metric and by expert judgment, over the 178 problems. The best value per column is in bold.

BEq+ produce rankings that differ substantially from SA-PASS$_{\text{soft}}$ (Figure 7), with the compile rate agreeing at only $\rho = 0.41$ and BEq+ passing 12.4%. This result over many independent participants is consistent with the metric-validity gap we measure against expert judgment in §6.1.

The best submissions use agentic pipelines that decompose each goal into auxiliary lemmas, and we observe the same trend in our experiments, where Claude Code (Opus 4.8) with Numina-Lean-Agent achieves the best result. We examine two factors, agentic behavior and decomposition. We label a submission as agentic when it runs more than three turns, and we measure decomposition by the number of declarations, such as lemma, def, or theorem. The verified score correlates with the number of declarations at Spearman's ρ of 0.59 and with the number of turns at 0.56, emphasizing that SHADOWBENCH requires multi-turn solving through decomposition. 59.7% of submissions add at least one auxiliary lemma or definition (3.6 on average), which is close to the 4.5 in the reference proofs (Table 9). Table 23 lists the top ten participants and their submission profiles.

H Annotation Interface

The 178 problems were annotated through a custom browser-based tool. The tool exposes three

Rank	SA-PASS$_{\text{soft}}$ %	Compile %	No Sorry %	Length (Char)	Auxiliary	Multiple declaration %
1	30.9	68.8	99.2	3,045	5.38	92.0
2	30.1	75.2	99.2	2,952	5.42	90.4
3	27.9	52.0	96.8	5,133	9.58	58.2
4	27.8	80.0	99.2	2,365	4.40	75.2
5	27.4	77.6	99.2	2,899	5.71	88.0
6	24.7	63.2	95.2	4,113	6.56	70.8
7	20.9	75.2	94.4	969	2.43	68.9
8	12.5	79.2	100.0	2,567	2.69	47.6
9	11.4	69.6	100.0	2,434	3.53	45.2
10	10.8	66.4	91.1	8,052	6.89	78.8

Table 23: Top ten teams on the SHADOWBENCH public challenge track, anonymized and ranked by SA-PASS$_{\text{soft}}$. Compile is the fraction of the 126 problems whose submitted solution compiles standalone. No Sorry counts the number of problems solved without sorry. Length indicates the average length of submission in characters. Auxiliary is the average number of auxiliary declaration per each problem solution beyond single canonical solution. Multiple declaration counts the number of solutions that used more than one declaration. The median SA-PASS$_{\text{soft}}$ across all 27 participants is 8.1%.

views, each captured below. Figure 8 is the main page, which shows the number of problems annotated for each area and level. Selecting a problem opens the detail page in Figure 9. The detail view renders the natural language statement and proof, and exposes metadata such as area, level, paper of origin, subfields, and type tags. This page also shows whether the problem is valid by compiling with the reference Lean code through the same Docker image used at evaluation time. To submit the new problem, curators fill out the form in Figure 10. This page accepts paired natural language text and Lean code, optional auxiliary imports, reference code that the examinee must complete, and hidden shadow checks used to evaluate semantic correctness.

I Prompt Templates

Every prompt used in this paper is reproduced below verbatim, one per figure. Placeholders of the form `{{ ... }}` are filled per problem by the serving pipeline. Raw text sources live under `prompts/` and `scripts/judge_prompt.py` in the code repository. The structured environments here mirror the `system`, `user`, and `model` roles in the API call.

Non-agentic prompt. We use the prompt shown in Figure 11 for non-agentic models.

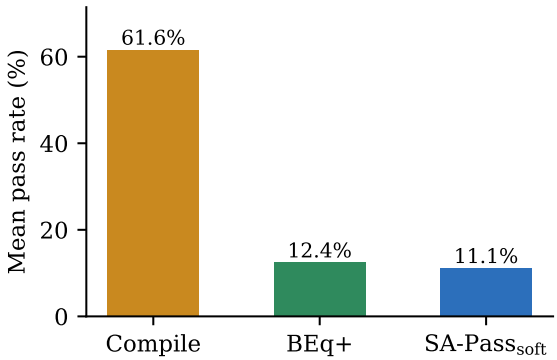
Agentic baseline. Figure 12 shows the direct agentic prompt used without Numina-Lean-Agent. The Claude Code rows invoke this prompt through the standard Claude Code CLI on the target Lean file. Codex uses the same task instruction in its own wrapper.

Agentic setting (Numina-Lean-Agent). Figure 13 is used for the + *Numina* rows of Table 3. Numina-Lean-Agent runs the coordinator prompt below, which instructs the agent to select one target `sorry` per session, sketch incrementally, and verify via `lean_diagnostic_messages`.

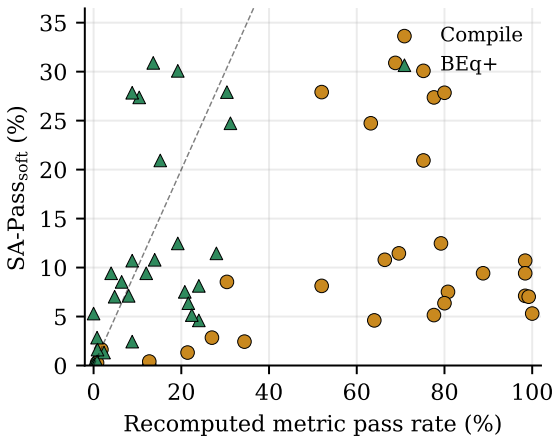
I.1 LLM-as-Judge Prompt

The judge prompt drives the Judge column in Table 12. For each candidate the builder fills `{{ informal_block }}` with the problem's NL statement (and NL proof, when the ablation condition provides one), `{{ formal_ref }}` with the reference Lean 4 statement with `sorry`, and `{{ candidate }}` with the model's generated Lean 4 theorem plus its proof.

The response parser accepts bare JSON, fenced JSON, and a permissive regex fallback. Verdicts are normalised to {`correct`, `incorrect`, `uncertain`}. All three are recorded per vote, then reduced by majority.



(a) Mean pass rate under compilation, BEq+, and SA-PASS$_{\text{soft}}$.



(b) Per participant, the recomputed metric pass rate against SA-PASS$_{\text{soft}}$.

Figure 7: Metric disagreement on the public challenge. In (a), compilation accepts $5.5\times$ as many outputs on average compared to SA-PASS$_{\text{soft}}$. In (b), points for compilation lie far from the diagonal, combining a high compile rate with a low SA-PASS$_{\text{soft}}$ score, which indicates over-acceptance.

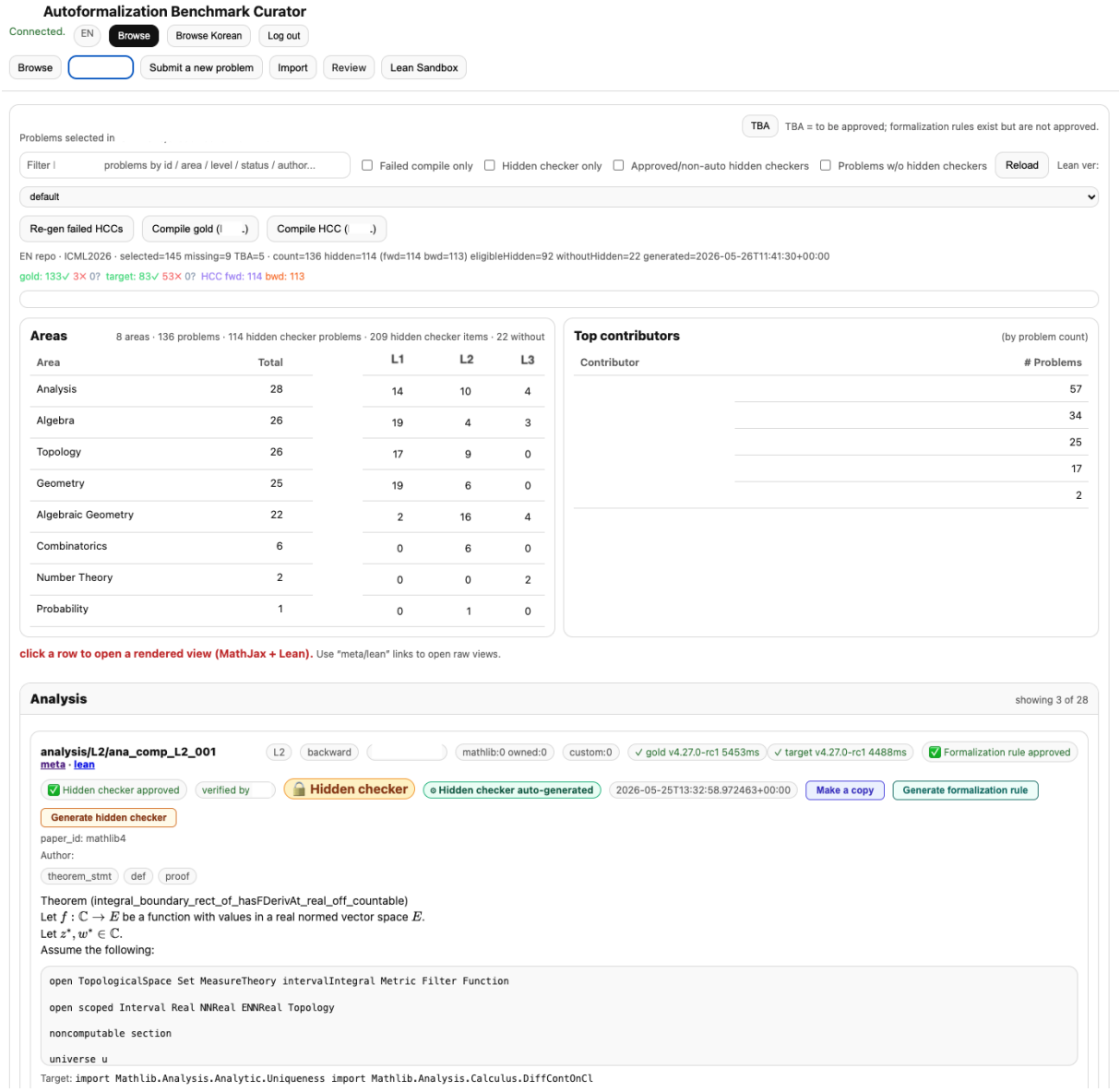


Figure 8: Data annotation landing page. The left panel summarises area by level coverage. The right panel ranks contributors by the number of problems they have annotated. Selecting a row in the left panel opens the problem detail in Figure 9.

algebra/L3/alg_grob_L3_001

Open meta | Open lean | Copy text | Copy lean | Copy meta.json | Edit | Move/Rename | Compile | ✓ Compiled successfully (321 ms) at 22/04/2026, 01:25:36

Compile target | ✓ Target compiled (284 ms) at 22/04/2026, 01:25:34 | Approve target | Hidden checker | Delete… | Close

ID	algebra/L3/alg_grob_L3_001
Area / Level	Algebra L3
paper_id	EGA
paper_title	
Author	
Types	theorem_stmt
Subfields	grobner-basis
Toolchain	—
Custom module	—
Created	2026-01-07T09:45:56+00:00

Math / NL text

Definition (Finite morphism)

A morphism of schemes $f : X \to Y$ is called \emph{finite} if it is affine and for every affine open subset $U = \mathrm{Spec}(A) \subset Y$, the preimage $f^{-1}(U)$ is affine, say $f^{-1}(U) = \mathrm{Spec}(B)$, where B is a finite A-module.

Definition (Projective morphism)

A morphism of schemes $f : X \to Y$ is called \emph{projective} if there exists an integer $n \geq 0$ and a closed immersion

$$i : X \hookrightarrow \mathbb{P}^n_Y$$

such that f factors as

$$X \xrightarrow{i} \mathbb{P}^n_Y \to Y,$$

where $\mathbb{P}^n_Y$ denotes the projective n-space over Y.

Theorem

Let $f : X \to Y$ be a finite morphism of schemes. Then f is projective.

Proof

Figure 9: Problem detail view (`algebra/L3/alg_grob_L3_001`). Header buttons surface the underlying `meta.json` and `gold.lean`, and the "Compile" / "Compile target" actions rerun the Lean type checker through the evaluation Docker image. The body renders the LaTeX natural language statement and proof.

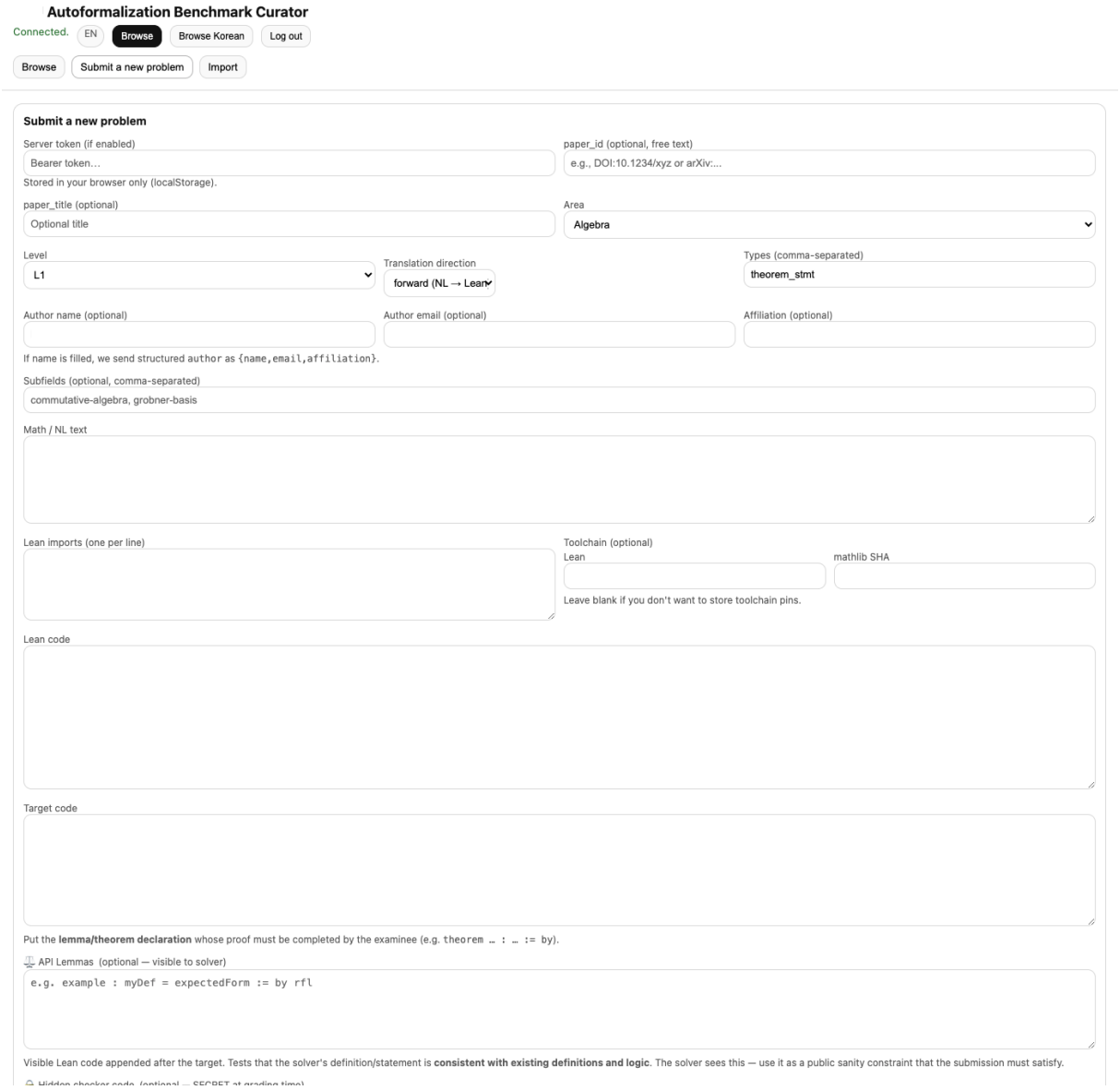


Figure 10: Submission form for a new problem. The form pairs natural language text (`Math / NL text`) with Lean code, accepts optional auxiliary lemma imports, and separates the public target code from an optional hidden checker block used for semantic evaluation.

System Prompt

You are a helpful assistant that can interact multiple times with a computer shell to solve programming tasks. However, for this specific interaction, you are acting as an expert Lean 4 mathematician tasked with implementing a formal Lean 4 declaration (theorem or definition) based on the provided context.
Your response must contain exactly ONE lean code block.
Include a THOUGHT section before your code where you explain your reasoning process. Format your response as shown in `<format_example>`.
`<format_example>`

Natural Language

```
THOUGHT: Your reasoning and analysis here, explaining how you deduce the
    implementation from the context.
```

Lean 4

```
your_formal_lean_code_here
```

`</format_example>`
Failure to follow these rules will cause your response to be rejected.

User Prompt

`<task_description>`
Your task is to provide the full formal Lean 4 code for the specific declaration named `{{ target_code_name }}`. The declaration is part of a larger code sequence, and you must implement it to fit seamlessly into the provided context.
The Code Context is provided below.

- `Header`: Imports and open namespaces.
- `Code Before`: Code appearing before the target.
- `Target Code Name`: The name of the definition or theorem you need to implement.
- `Natural Language Description`: A description of the target code in natural language.
- `Code After`: Code appearing after the target.

Header:

Lean 4

```
{{ header }}
```

Code Before:

Lean 4

```
{{ before_target_code }}
```

Target Code Name:

Natural Language

```
{{target_code_name}}
```

Natural Language Description:

Natural Language

```
{{target_natural_language}}
```

Code After:

Lean 4

```
{{ after_target_code }}
```

<instructions>
 # Task Instructions
Overview
You are an expert Lean 4 mathematician. You must provide the complete and correct Lean 4 implementation for `{{ target_code_name }}`.

- If it is a theorem, provide the statement and the full proof (do not use `sorry`).
- If it is a definition, provide the full definition.
- Pay close attention to `Code Before` and `Code After` to infer the correct type signatures, variable names, and logical dependencies.
- Ensure your code compiles and integrates correctly with the surrounding context.

Output Coding Rules
1. A **THOUGHT** section where you explain your reasoning:

- Analyze the `Code Before` and `Code After` to deduce the purpose and signature of `{{ target_code_name }}`.
- Plan the implementation.

2. A single lean code block with your implementation.
Format your responses like this:
<format_example>

Natural Language

```
THOUGHT: The context shows a function \texttt{foo} is used to add two numbers.
I will implement \texttt{foo} as a definition taking two Nats and returning
    their sum.
```

Lean 4

```
def foo (n m : Nat) : Nat := n + m
```

</format_example>
CRITICAL REQUIREMENTS:

- Your response SHOULD include a THOUGHT section.
- Your response MUST include EXACTLY ONE lean block.
- This block MUST contain the COMPLETELY NEW implementation for `{{ target_code_name }}`.
- Do NOT repeat the `Header` or `Code Before` or `Code After`. Only output the code for `{{ target_code_name }}`.

</instructions>

AI Response

Natural Language

```
{{thoughts}}
```

Lean 4

```
{{ formal_proof }}
```

Figure 11: Prompt template for autoformalizing to Lean 4 code.

User Prompt

Complete the target Lean theorem file, making it sorry-free and ensuring it compiles without errors. Use the tool `lean_diagnostic_messages` to verify the file. Errors mean "severity 1" in the response.
IMPORTANT:

- Do NOT write a detailed full proof plan upfront. Instead, write a one or two sentence high-level idea, then implement one step at a time, checking `lean_goal` after each step to plan the next.
- Verify the file again after each update using `lean_diagnostic_messages`.
- You may add helper lemmas, but do not create new axioms.

Tips:

- Use `simp` first, then `simp?` to get minimal simp lemmas.
- Use `native_decide` for computational results.
- Use `#eval` to evaluate expressions.
- Use `norm_cast` for type conversions.
- Use `apply?` to find applicable lemmas.
- If `decide` times out, do NOT simply increase `maxHeartbeats`. Instead, write a symbolic proof using mathematical reasoning and lemmas from Mathlib.

At the end of your response, include:

Natural Language

```
END_REASON:{reason}
```

where `{reason}` is:

- `LIMIT`: stopped due to limits or there are still sorries/errors.
- `COMPLETE`: the file is sorry-free AND compiles without errors.

IMPORTANT: Use `lean_diagnostic_messages` to verify, do not use `lake build`.

Figure 12: Prompt template for the direct agentic baseline without the retrieval harness.

User Prompt

Please analyze all sorries in the file and choose ONLY ONE that you believe is most approachable or strategically important to work on.

 # Your task: write Lean formal proofs

Your task is to write Lean 4 formal proofs that compile successfully. Replace `sorry` with executable Lean code (tactics, term-mode proofs, etc.) that passes the Lean compiler. Choose ONLY ONE target lemma/sorry for this entire session, focus on it and try your best to complete it.

 # [Critical prohibition] no natural language proofs

1. ZERO TOLERANCE FOR NARRATIVE COMMENTS. You are forbidden from using comments (`/- ...-/` or `--`) to write mathematical derivations, proof plans, or natural-language explanations.
2. STRICT COMMENT LENGTH LIMIT. No single comment block may exceed 42 lines.
3. CODE IS THE EXPLANATION. If a logic step is complex enough to need an explanation, extract it into a new helper lemma.
4. CONSTRAINT COMMENT. Prohibit sequences of 5 or more consecutive comment blocks unless interleaved with valid Lean code.

 # Session focus rules

At session start:

1. Identify all remaining sorries or failed proofs.
2. Select EXACTLY ONE target that is most approachable and strategically important.
3. Announce: `"TARGET FOR THIS SESSION: [lemma_name] at line [line_number]"`.

During session: only work on your selected target and any new helper lemmas you create for it.
End session with ONE of:

- `SELECTED_TARGET_COMPLETE`: target fully proven.
- `COMPLETE`: all sorries in folder proven.
- `LIMIT`: stopped due to token/time/error limits.

 # Incremental planning
Do NOT write a detailed upfront plan. Instead:

1. Abstract sketch: a one or two sentence high-level idea.
2. One step at a time. After each step, use `lean_goal` and `lean_diagnostic_messages`.
3. Re-plan from the actual proof state, not from your initial sketch.

 # Tools

- `lean_diagnostic_messages`, `lean_goal`, `lean_leandex`, `gemini_informal_prover`, `create_formal_sketch`, `discussion_partner`.
- Do NOT use `lake build` or `lean_build`. Always use `lean_diagnostic_messages`.

 # End format
At the very end of the response include exactly one line:

Natural Language

```
END_REASON:{reason}
```

Figure 13: Prompt template for the Numina-Lean-Agent harness used with Claude Code, Codex, and Qwen3 235B.

System Prompt

You are an expert in Lean 4 and Mathlib. Decide whether a candidate Lean 4 statement+proof semantically establishes the INTENDED mathematical theorem shown alongside it.
Return ONE JSON object and nothing else:

Natural Language

```
{"verdict": "correct" | "incorrect" | "uncertain",
 "reason":  "<1-3 sentences>"}
```

- `"correct"`: (a) the candidate's stated theorem is mathematically equivalent to the intended claim, AND (b) its proof actually justifies that claim.
- `"incorrect"`: statement weaker/stronger/wrong OR proof unsound, circular, or relies on a broken auxiliary.
- `"uncertain"`: artifact truncated or unreadable.

User Prompt

 # Intended problem (informal)

Natural Language

```
{{ informal_block }}
```

 # Intended formal statement (reference Lean 4, proof body is `sorry`)

Lean 4

```
{{ formal_ref }}
```

 # Candidate (Lean 4, statement + attempted proof)

Lean 4

```
{{ candidate }}
```

Return your JSON verdict now.

AI Response

Natural Language

```
{"verdict": "{{ verdict }}",
 "reason":  "{{ reason }}"}
```

Figure 14: Prompt template for the LLM-as-judge proof correctness evaluator.

User Prompt

```
You are generating ABM hidden checker code for Lean 4.

## Task
The natural-language text below states a mathematical result.  It contains
several distinct sub-claims (e.g. conjunctions, consequences, specialisations,
unfolded definitions, or other properties implied by the main statement).

Step 1 -- identify the sub-claims present in the NL text.
Step 2 -- for each target declaration listed below, write one Lean
  example per sub-claim that PROVES that sub-claim follows from the solver's
  declaration.  The checker is appended after the solver's output, so the
  solver's declaration name is already in scope.

## Rules for every generated checker item
- It must be a STRICT sub-statement of the corresponding target item: a
  projection (.1/.2/.mp/.mpr), a specialisation (∀ instantiated at a concrete
  value), an unfolded/simp-normal form, an iff direction, or a direct
  consequence that the solver must have actually proved.
- It must reference and use the target declaration name in its proof.
- It must NOT restate the full target declaration and prove it by exact/simpa
  using the target declaration.  That is just copying the theorem, not a hidden
  checker.
- It must NOT copy from gold.lean proofs. You are given only target declaration
  headers and text matches; generate small shadows from those declarations.
- It must contain no sorry, admit, axiom, constant, namespace, or end.
- Prefer anonymous top-level `example` declarations.  Do not introduce named
  declarations unless needed.
- Use explicit binders; do not assume variables that are not in scope.  The
  declaration's parameters are ONLY what appears in its signature before the
  `:` (the return type).  Nothing after `:=` is a parameter.
- An `↔` (Iff) proposition is NOT a function and cannot be applied to
  arguments.  Never write `exact (iff_lemma arg1 arg2)` or `simpa using
  (iff_lemma arg1 arg2)`.  Use `.mp`, `.mpr`, `.1`, or `.2` to access
  its directions: `exact (iff_lemma arg1 arg2).mp`.
- For conjunction conclusions, generate one shadow per conjunct using `.1`,
  `.2`, `.2.1`, etc.  Do not add helper-only or completeness-certificate items:
```

```
  every checker declaration is scored independently and must use the target
  declaration in its proof body.
- For iff conclusions, generate `.mp` and `.mpr` shadows.  Do not add a
  completeness certificate.

## Proof robustness requirement
Solvers may declare the same theorem with a different binder style (making some
arguments implicit, reordering type-class parameters, etc.) while being
mathematically correct.  To maximise compatibility, every checker item MUST use
a `first |...|..` tactic block with the following three alternatives in order:

1. Direct projection using the gold argument list:
   `exact (TARGET ARGS).PROJECTION`
2. Let-bound variant (separates elaboration, helps with universe/instance issues
    ):
   `exact (let _hc := TARGET ARGS; _hc.PROJECTION)`
3. Argument-agnostic fallback -- lets Lean's unifier determine how to apply the
   solver's declaration regardless of how the solver declared its binders:
   `suffices _hc : FULL_CONCLUSION by exact _hc.PROJECTION; apply TARGET <;>
    assumption`

Example for a conjunction conjunct (`.1` of `foo (n : ℕ) (h : n > 0) : A n ∧ B n
    `):
```lean
example (n : ℕ) (h : n > 0) :
    A n := by
  first
  | exact (foo n h).1
  | exact (let _hc := foo n h; _hc.1)
  | (suffices _hc : A n ∧ B n by exact _hc.1; apply foo <;> assumption)
```

Example for an iff forward direction (`bar (x : α) : P x ↔ Q x`):
```lean
example (x : α) :
    P x → Q x := by
  first
  | exact (bar x).mp
  | exact (bar x).1
  | exact (let _hc := bar x; _hc.mp)
  | (suffices _hc : P x ↔ Q x by exact _hc.mp; apply bar <;> assumption)
```

For non-structural sub-claims (specialisations, unfoldings, consequences), apply
     the
same pattern: lead with the natural `exact`, follow with a `let`-bound variant,
    and
close with the `apply TARGET <;> assumption` fallback wrapped in `suffices`.

Problem
Problem id: {problem_id}
Target declarations (must each appear in at least one checker): {target_names}
Target source: {target_source_label}

Target items:
{target items}

Formalization rules
{formalization rules}

Natural-language text (source of sub-claims)
{natural-language text}

Output format
If multiple checker declarations are alternative versions of the same target
    substatement,
give them identical target/substatement metadata; their Lean doc-comment labels
    will be
```

```
normalized to Shadow 1-1, Shadow 1-2, etc. Distinct substatements remain Shadow
    1,
Shadow 2, etc.

Return ONLY valid JSON -- no markdown, no prose, no extra keys:
{
  "checker_code": "<complete Lean code block to append after solver output>",
  "items": [
    {
      "target_index": 1,
      "target_name": "foo",
      "checker_names": ["hidden_foo_part1"],
      "method": "projection/specialisation/unfolding/iff_direction/consequence",
      "substatement": "one-line description of the sub-claim"
    }
  ],
  "rationale": "one-line summary"
}
```

Figure 15: Prompt for generating forward checker theorems, i.e. strict sub-statements (shadows) S_i with $T \Rightarrow S_i$, from the reference formal statement. Deterministic shadows are emitted first and this prompt fills the remaining cases.

User Prompt

```
You are proving a SUFFICIENCY certificate in Lean 4: that a set of sub-
    statements JOINTLY imply a gold conclusion.

Gold target `{target_name}` has conclusion:
    {conclusion}

You are given these sub-statements as hypotheses:
{shadow hypotheses}

Write a single Lean 4 `example` that takes exactly those hypotheses and proves
    the gold conclusion:

example{binder prefix}
{shadow hypotheses} :
    {conclusion} := by
  <proof>

Rules:
- Use ONLY the given hypotheses h_bc_i (do not reference the original target or
    any solver theorem).
- Prefer `exact ⟨...⟩` for conjunctions, `⟨.mp, .mpr⟩` style for iff, or `
    constructor`/`refine`/`tauto`.
- Output ONLY the Lean code block, no prose.
```

Figure 16: Prompt for the backward sufficiency certificate $\bigwedge_i S_i \Rightarrow T$, used as a fallback when the certificate cannot be produced structurally.

J Shadow Check Case Studies

These case studies instantiate the forward checks of §3.2 on real statements. In the first case (Appendix J.1), the intended statement T says a finite morphism is projective, that is, finite $\Rightarrow$ projective. Its shadows (Equation (2)) are affine (S_1) and `HasProjectiveEmbedding` (S_2), so the forward checks $T \Rightarrow S_i$ are finite $\Rightarrow$ affine and finite $\Rightarrow$ `HasProjectiveEmbedding`, written as the two Lean examples that follow. The separation condition (Equation (4)) keeps the two shadows distinct, so passing the forward checks is not vacuous. This non-triviality is what a complete shadow set requires, and the backward check $S_1 \wedge \cdots \wedge S_n \Rightarrow T$ is the completeness half it protects. Each later case names its own target T and the shadows S_i that must follow from it.

This appendix walks through eight case studies that illustrate how these shadow checks apply to real mathematical statements. The first case is worked out in detail. The others are sketched.

J.1 Algebraic Geometry: Finite Morphisms Are Projective

Consider the textbook theorem that finite morphisms are projective. Let

$$\begin{aligned} A &= \text{finite morphism}, \\ B &= \text{projective morphism} \end{aligned} \tag{1}$$

Write $\texttt{HasProjectiveEmbedding}(f)$ for the property that X admits a closed S-immersion into $\mathbb{P}^n_S$ for some $n \geq 0$ (where S is the target of f). Natural shadows are

$$\begin{aligned} A' &= \text{affine morphism}, \\ B' &= \texttt{HasProjectiveEmbedding} \end{aligned} \tag{2}$$

The forward checks rely on the implications

$$\begin{aligned} \text{finite} &\Rightarrow \text{affine}, \\ \text{projective} &\Rightarrow \texttt{HasProjectiveEmbedding}. \end{aligned} \tag{3}$$

The second holds because a projective morphism $f : X \to S$ factors as a closed S-immersion $X \hookrightarrow \mathbb{P}^n_S$ followed by the projection, so X admits the required closed S-immersion by construction. The separation condition holds because an affine morphism need not have property `HasProjectiveEmbedding`: a closed S-subscheme of $\mathbb{P}^n_S$ is proper over S, but affine morphisms need not be proper. Over a field k, the structure morphism

$$\mathbb{A}^1_k \to \operatorname{Spec} k$$

is affine, but $\mathbb{A}^1_k$ admits no closed immersion into any $\mathbb{P}^n_k$, since any such subscheme would be proper. Thus

$$A' \not\Rightarrow B'. \tag{4}$$

The public task asks the solver to define finite and projective morphisms and prove that finite implies projective. The hidden checker contains the forward-check one-liners:

```
example
{X Y : Scheme}
{f : X ⟶ Y}
(hf : IsFinite f) :
IsAffineHom f := by
  simpa using IsFinite.isAffine hf
```

and the check for the second shadow, discharged by applying the solver's theorem

```
example
{X S : Scheme}
{f : X ⟶ S}
(hf : IsFinite f) :
HasProjectiveEmbedding f := by
  simpa [HasProjectiveEmbedding]
    using finite_implies_projective hf
```

If the first check fails, the LLM fallback attempts to prove `IsFinite` $\Rightarrow$ `IsAffineHom` from the solver's definition. The second check is not subject to fallback, since it must be discharged through the solver's theorem.

J.2 Projective Products and Hidden Properness

A related theorem is that the fiber product of two projective schemes over a base is projective, classically via the Segre embedding. The public theorem may be:

$$\begin{aligned} X \to S \text{ projective}, Y \to S \text{ projective} \\ \Rightarrow X \times_S Y \to S \text{ projective} \end{aligned} \tag{5}$$

A hidden forward check can ask whether the product morphism has property `HasProjectiveEmbedding`, i.e., whether $X \times_S Y$ admits a closed S-immersion into some $\mathbb{P}^n_S$. This tests that the product theorem returns projectivity in a usable form and that the solver's definition of projectivity connects to the trusted `HasProjectiveEmbedding` predicate. Here

the target T is the projectivity of the product morphism, and the shadow S_1 is the property that the product has `HasProjectiveEmbedding`, so the forward check $T \Rightarrow S_1$ asks the product to admit a closed S-immersion into some $\mathbb{P}^n_S$.

J.3 Compiled False Positive: Projective as Proper

Claude Opus 4.6 and Claude Sonnet 4.6 both produced the following definition for an algebraic-geometry problem about projective morphisms. The submission compiles, but it defines projectivity as properness:

```
class IsProjective
{S X : Scheme}
(f : X ⟶ S) : Prop where
  isProper : IsProper f
```

With this definition, the generated theorem is immediate:

```
import Mathlib

namespace ABM
namespace algebraic_geometry
namespace L3
namespace alg_sche_L3_003

open CategoryTheory AlgebraicGeometry

class IsProjective
{S X : Scheme}
(f : X ⟶ S) :
Prop where
  isProper : IsProper f

open CategoryTheory.Limits

theorem projective_isProper
{S X : Scheme}
(f : X ⟶ S) [IsProjective f] :
IsProper f :=
  IsProjective.isProper

end alg_sche_L3_003
end L3
end algebraic_geometry
end ABM
```

Compile rate marks this candidate as correct because the theorem block type-checks. SA-PASS rejects it: the hidden shadow checks require the generated notion of projectivity to support the intended algebraic-geometric content, not only the weaker properness property. Here the shadow S is again `HasProjectiveEmbedding`, and the forward check $\widehat{T} \Rightarrow S$ fails, because the generated statement proves only properness, which does not imply `HasProjectiveEmbedding`.

J.4 Compiled Weakening: Complex Structure Dropped

The forward check verifies the weakened statements. Claude Code (Opus 4.8) with Numina-Lean-Agent produced the following statement for a Fourier analysis problem, whose intended lemma is a Gaussian truncation limit used toward Fourier inversion. The intended statement is complex valued:

```
-- Intended statement (abridged)
variable {E : Type*}
  [NormedAddCommGroup E]
  [NormedSpace ℂ E] {f : V → E}

lemma tendsto_integral_cexp_sq_smul
    (hf : Integrable f) :
    Tendsto
      (fun c : ℝ => ∫ v : V,
        cexp (- c⁻¹ * ‖v‖^2) • f v)
      atTop (nhds (∫ v : V, f v))
```

The generated output compiles with a correct proof of a real-valued variant. It replaces `cexp` and the $\mathbb{C}$ module structure on E with `Real.exp` and an $\mathbb{R}$ module structure, and generalizes the volume measure to an arbitrary measure:

```
-- Submission (abridged)
{E : Type*} [NormedAddCommGroup E]
  [NormedSpace ℝ E] {f : V → E}
                    {μ : Measure V}

lemma tendsto_integral_cexp_sq_smul
    (hf : Integrable f μ) :
    Tendsto
      (fun c : ℝ => ∫ v,
        Real.exp
          (- c⁻¹ * ‖v‖^2) • f v ∂μ)
      atTop (nhds (∫ v, f v ∂μ))
```

The backward check passes, since the shadows derived from the complex statement imply the real variant. The forward check fails, since the real variant does not recover the complex shadow. Compile rate considers this generation as correct, and SA-PASS rejects it as a weakening.

J.5 Compiled False Positive: Conclusion Assumed as a Hypothesis

A compiling submission can also prove a different statement. Claude Code (Opus 4.8) with Numina-Lean-Agent produced the following statement for Brahmagupta's formula (`geo_gen_L2_007`). The intended theorem computes the area of a cyclic quadrilateral, stated as the measure of the convex hull of the four vertices:

```
-- Intended statement (abridged)
theorem brahmagupta_formula
    {A B C D : EuclideanSpace ℝ(Fin 2)}
    (h_cyclic : Concyclic
        ({A, B, C, D} : Set _))
    (h_convex :
        (openSegment ℝ A C
        ∩ openSegment ℝ B D).Nonempty):
    let s := (dist A B + dist B C
        + dist C D + dist D A) / 2
    let K := (volume (convexHull ℝ
        ({A, B, C, D} : Set _))).toReal
    K = sqrt ((s - dist A B)
        * (s - dist B C)
        * (s - dist C D)
        * (s - dist D A))
```

The submission compiles with a correct proof of a different statement. It receives the value K as a hypothesis that already encodes the triangle decomposition of the area, and replaces concyclicity of the four points with an angle condition:

```
-- Submission (abridged)
theorem brahmagupta_formula
    (A B C D : P) (a b c d s K : ℝ)
    (ha : a = dist A B)
    (hb : b = dist B C)
    (hc : c = dist C D)
    (hd : d = dist D A)
    (hs : s = (a + b + c + d) / 2)
    (hcyclic :
        EuclideanGeometry.angle D A B
        + EuclideanGeometry.angle
        B C D = π)
    (harea : K = 1 / 2 * a * d
        * Real.sin
          (EuclideanGeometry.angle
            D A B)
        + 1 / 2 * b * c
        * Real.sin
          (EuclideanGeometry.angle
            B C D))
    (hK : 0 ≤ K) :
    K = Real.sqrt ((s - a) * (s - b)
        * (s - c) * (s - d))
```

The resulting statement is an identity about the assumed value K, not a statement about the area. Both check directions fail, since the submission neither implies nor is implied by the shadows of the intended theorem. Compile rate counts this submission as correct, and SA-PASS rejects it.

J.6 Backup Declarations on ProofNet

The larger models often attach a backup declaration to a long proof attempt. Claude Opus 4.6 produced the following output for Herstein Exercise 2.1.18 on ProofNet, which asks for an element $a \neq 1$ with $a = a^{-1}$ in a group of even order. The generated statement matches the intended one, and the output carries a second declaration that ends in `sorry`:

```
-- Claude Opus 4.6 (abridged)
theorem exercise_2_1_18
  {G : Type*} [Group G]
  [Fintype G] (hG2 : Even (card G)) :
  ∃ (a : G), a ≠ 1 ∧ a = a⁻¹ := by
  have h : ∃ a : G, a ≠ 1 ∧ a * a = 1
  := by
    ... -- long proof attempt

theorem exercise_2_1_18'
  {G : Type*} [Group G]
  [Fintype G] (hG2 : Even (card G)) :
  ∃ (a : G), a ≠ 1 ∧ a = a⁻¹ := by
  sorry
```

The evaluation takes the last theorem block as the candidate (§5.1), which here is the backup declaration, so the output is rejected. Claude Haiku 4.5 emits a single declaration with the same statement for this problem and passes.

J.7 Flat Morphisms and Openness

For flat morphisms, a classical theorem states that a flat morphism locally of finite presentation is open. A benchmark can use

$$\begin{aligned} A &= \text{flat and locally of finite presentation,} \\ B &= \text{open map.} \end{aligned} \tag{6}$$

Possible shadows include generalizing maps and universally open morphisms, depending on the available benchmark environment. Here the target T is $A \Rightarrow B$, and the shadows S_i are properties such as being a generalizing map or a universally open morphism that B should entail, so each forward check $T \Rightarrow S_i$ tests one consequence of openness. Here the separation condition must be checked carefully: if the chosen shadow of A already implies the chosen shadow of B in that environment, then the item is vulnerable to collapse. The framework forces the benchmark designer to make this relationship explicit before releasing the task.

J.8 Homotopy and Homology

In algebraic topology, many one step facts are already explicit in mature libraries: path homotopy is an equivalence relation, path homotopy respects concatenation, homotopy equivalences compose, and so on. Such facts should not be visible target theorems if the goal is to evaluate autoformalization beyond retrieval.

The shadow based evaluation suggests using derived tasks instead. For instance, a public theorem may ask for a bundled threefold product statement for homotopy equivalences, while hidden checks specialize it to product with an identity factor. In homology, a public theorem may bundle the degree zero homology computation of totally disconnected spaces with positive degree vanishing, while hidden checks specialize to concrete degrees. Here each public target T is a bundled statement, and the shadows S_i are its specializations, so the forward checks $T \Rightarrow S_i$ project the bundle onto one component in the bundled-conclusion pattern of §3.2. These tasks are not merely lookups of named declarations. They require composing library facts into reusable results.

K Complete Shadow Set Checker Patterns

K.1 Bundled Conclusions

For a bundled theorem, the hidden checker can compile each projection.

```
-- Expected submission:
-- theorem main : A -> B1 /\ B2 /\ B3

example (hA : A) : B1 := by
  exact (main hA).1

example (hA : A) : B2 := by
  exact (main hA).2.1

example (hA : A) : B3 := by
  exact (main hA).2.2
```

K.2 Equality as Two Inequalities

For equality-valued theorems, the hidden checker can compile both inequality consequences.

```
-- Expected submission:
-- theorem main (hA : A) : f B = f C

example (hA : A) : f B <= f C := by
  exact le_of_eq (main hA)

example (hA : A) : f C <= f B := by
  exact ge_of_eq (main hA)
```

The exact lemmas depend on the proof assistant and ordered structure. The benchmark uses the idiomatic library lemmas for converting equality into the two inequality directions.

K.3 Saddle-Point Derivatives

This example on the next page shows a complete shadow set for a theorem whose conclusion is a pair of first-order conditions. The intended conclusion is $Dx = 0 \wedge Dz = 0$. The hidden checker splits it into two shadows and then checks that the two shadows recover the full conclusion.

```
/-- Shadow 1: the derivative of the `x`-section vanishes. -/
example
    {n m : ℕ}
    {f : (Fin n → ℝ) × (Fin m → ℝ) → ℝ}
    {x0 : Fin n → ℝ} {z0 : Fin m → ℝ}
    {Dx : (Fin n → ℝ) →L[ℝ] ℝ}
    {Dz : (Fin m → ℝ) →L[ℝ] ℝ}
    (hs : ∀ x z, f (x0, z) ≤ f (x0, z0) ∧ f (x0, z0) ≤ f (x, z0))
    (hx : HasFDerivAt (fun x => f (x, z0)) Dx x0)
    (hz : HasFDerivAt (fun z => f (x0, z)) Dz z0) :
    Dx = 0 := by
  exact (saddle_sections_hasFDerivAt_eq_zero hs hx hz).1

/-- Shadow 2: the derivative of the `z`-section vanishes. -/
example
    {n m : ℕ}
    {f : (Fin n → ℝ) × (Fin m → ℝ) → ℝ}
    {x0 : Fin n → ℝ} {z0 : Fin m → ℝ}
    {Dx : (Fin n → ℝ) →L[ℝ] ℝ}
    {Dz : (Fin m → ℝ) →L[ℝ] ℝ}
    (hs : ∀ x z, f (x0, z) ≤ f (x0, z0) ∧ f (x0, z0) ≤ f (x, z0))
    (hx : HasFDerivAt (fun x => f (x, z0)) Dx x0)
    (hz : HasFDerivAt (fun z => f (x0, z)) Dz z0) :
    Dz = 0 := by
  exact (saddle_sections_hasFDerivAt_eq_zero hs hx hz).2

/-- Completeness certificate: Shadow 1 ∧ Shadow 2 ⇒ Dx = 0 ∧ Dz = 0. -/
example
    (hDx :
      ∀ {n m : ℕ}
        {f : (Fin n → ℝ) × (Fin m → ℝ) → ℝ}
        {x0 : Fin n → ℝ} {z0 : Fin m → ℝ}
        {Dx : (Fin n → ℝ) →L[ℝ] ℝ}
        {Dz : (Fin m → ℝ) →L[ℝ] ℝ},
        (∀ x z, f (x0, z) ≤ f (x0, z0) ∧ f (x0, z0) ≤ f (x, z0)) →
        HasFDerivAt (fun x => f (x, z0)) Dx x0 →
        HasFDerivAt (fun z => f (x0, z)) Dz z0 →
        Dx = 0)
    (hDz :
      ∀ {n m : ℕ}
        {f : (Fin n → ℝ) × (Fin m → ℝ) → ℝ}
        {x0 : Fin n → ℝ} {z0 : Fin m → ℝ}
        {Dx : (Fin n → ℝ) →L[ℝ] ℝ}
        {Dz : (Fin m → ℝ) →L[ℝ] ℝ},
        (∀ x z, f (x0, z) ≤ f (x0, z0) ∧ f (x0, z0) ≤ f (x, z0)) →
        HasFDerivAt (fun x => f (x, z0)) Dx x0 →
        HasFDerivAt (fun z => f (x0, z)) Dz z0 →
        Dz = 0) :
    ∀ {n m : ℕ}
      {f : (Fin n → ℝ) × (Fin m → ℝ) → ℝ}
      {x0 : Fin n → ℝ} {z0 : Fin m → ℝ}
      {Dx : (Fin n → ℝ) →L[ℝ] ℝ}
      {Dz : (Fin m → ℝ) →L[ℝ] ℝ},
      (∀ x z, f (x0, z) ≤ f (x0, z0) ∧ f (x0, z0) ≤ f (x, z0)) →
      HasFDerivAt (fun x => f (x, z0)) Dx x0 →
      HasFDerivAt (fun z => f (x0, z)) Dz z0 →
      Dx = 0 ∧ Dz = 0 := by
  intro n m f x0 z0 Dx Dz hs hx hz
  exact ⟨hDx hs hx hz, hDz hs hx hz⟩
```

L Author Contact Information

We list the contact information for all authors below.

- **Hojae Han**, Electronics and Telecommunications Research Institute. hojae.han@etri.re.kr
- **Jongyoon Kim**, Interdisciplinary Program in Artificial Intelligence, Seoul National University. john.jongyoon.kim@snu.ac.kr
- **Sanghyeok Park**, Department of Mathematical Sciences, Seoul National University. 202123018@snu.ac.kr
- **Dongwook Cheon**, Department of Mathematical Sciences, Seoul National University. dongwook0826@snu.ac.kr
- **Yeachan Park**, Department of Mathematics and Statistics, Sejong University. ychpark@sejong.ac.kr
- **Myung Jae Jeon**, Department of Mathematics, University of Maryland, College Park. mjjeon@umd.edu
- **Sunjong Choe**, Department of Mathematical Sciences, Seoul National University. sunjc@snu.ac.kr
- **Soonho Kong**, Amazon Web Services. soonho@amazon.com
- **Wonseok Hur**, Department of Computer Science and Engineering, Seoul National University. hws0728jik@snu.ac.kr
- **Seung-won Hwang**, Department of Computer Science and Engineering, Seoul National University. seungwonh@snu.ac.kr
- **Donghoon Hyeon**, Department of Mathematical Sciences, Seoul National University. dhyeon@snu.ac.kr