香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

*LA*anguage *R*easoning & *K*nowledge Lab

Toward Bridging the Informal-Formal Gap

Haocheng Wang

HKUST (GZ)

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

Motivation

FORMALRX: Rectify and eXamine Semantic Failures in Autoformalization

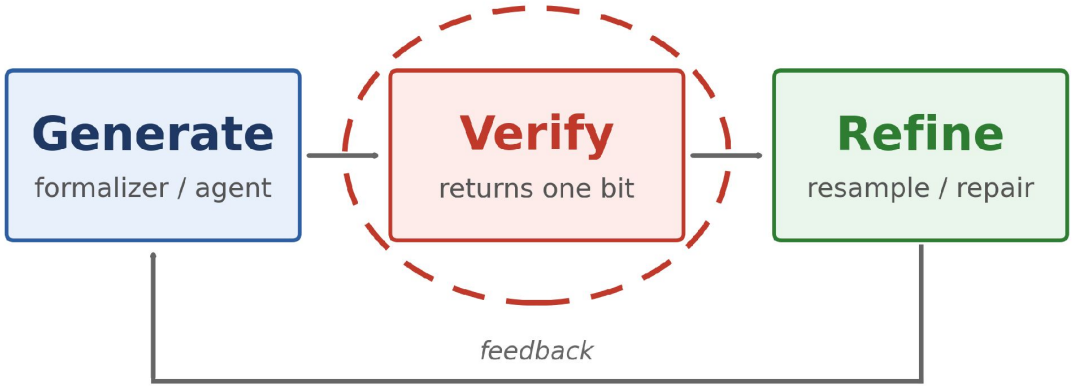
Haocheng Wang [*1 2] Baiyu Huang [*1] Yingjia Wan [*3] Xiao Zhu [1] Xiaoyang Liu [4]
Yinya Huang [† 5] Zhijiang Guo [† 1 6]

Autoformalization translates informal math into machine-checkable Lean code, but

😟A wrong translation is worse than no translation at all.

😟Existing evaluators only return a single binary verdict with no explanation.

😟An unexplained verdict cannot support human debugging or model improvement.



Scan for Project

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)
ETH zürich
UCLA
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY
ETH AI CENTER

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

TLDR

FORMALRX: Rectify and eXamine Semantic Failures in Autoformalization

Haocheng Wang [*1,2] Baiyu Huang [*1] Yingjia Wan [*3] Xiao Zhu [1] Xiaoyang Liu [4]
Yinya Huang [†5] Zhijiang Guo [†1,6]

1. Existing evaluations provide only opaque binary verdicts or scalar scores, offering no insight into where or why translations fail. ***FormalRx*** transforms this into actionable feedback: alignment verdict, error type, location, and correction.

2. At its core is the ***SCI Error Taxonomy***, a hierarchical scheme that decomposes autoformalization errors into 28 categories with strict priority ordering. Building on it, FormalRx provides four capabilities: alignment verdicts, error categorization, error localization, and correction.

3. We also release ***FormalRx-8B***, a diagnostic model trained on fine-grained annotations, and ***FormalRx-Test***, the first fine-grained diagnostic benchmark for autoformaliztaion.

📝 https://arxiv.org/abs/2607.04655

🤗 https://huggingface.co/collections/LARK-Lab/formalrx

📦 https://huggingface.co/datasets/LARK-Lab/FormalRx-Test

🔗 https://openreview.net/forum?id=toNy7vSxa3

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU) ETH zürich UCLA 上海交通大學 SHANGHAI JIAO TONG UNIVERSITY ETH AI CENTER

Scan for Project

Compiled, yet misaligned

✏️ Select a subset A from the set $\{1, 2, \cdots, 100\}$ such that there are no two consecutive integers in A. How many ways are there to do this? Show that it is $\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{102} - \left(\frac{1-\sqrt{5}}{2}\right)^{102}\right]$.

从集合 $\{1, 2, \cdots, 100\}$ 中选取一个子集 A, 使得 A 中不存在连续两个整数. 问有多少种选取方法? 答案是 $\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{102} - \left(\frac{1-\sqrt{5}}{2}\right)^{102}\right]$.

Misaligned：使用[0,100)而不是[1,100]

```
example :
  Finset.card (Finset.filter (fun A : Finset ℕ => ∀ i ∈ A, i + 1 ∉ A)
  (Finset.powerset (Finset.range 100))) =
  (1 / Real.sqrt 5) * (((1 + Real.sqrt 5) / 2) ^ 102 - ((1 - Real.sqrt 5) / 2) ^ 102) :=
```

Compiled, yet misaligned

✏️ Select a subset A from the set $\{1, 2, \cdots, 100\}$ such that there are no two consecutive integers in A. How many ways are there to do this? Show that it is $\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{102} - \left(\frac{1-\sqrt{5}}{2}\right)^{102}\right]$.

Misaligned：使用[0,100)而不是[1,100]

```
example :
  Finset.card (Finset.filter (fun A : Finset ℕ => ∀ i ∈ A, i + 1 ∉ A)
  (Finset.powerset (Finset.range 100))) =
  (1 / Real.sqrt 5) * (((1 + Real.sqrt 5) / 2) ^ 102 - ((1 - Real.sqrt 5) / 2) ^ 102)
:= by sorry
```

Aligned：使用与原问题精确对应的定义域[1,100]

```
example:
  Finset.card (Finset.filter (fun A : Finset ℕ => ∀ i ∈ A, i + 1 ∉ A)
  (Finset.powerset (Finset.Icc 1 100))) =
  (1 / Real.sqrt 5) * (((1 + Real.sqrt 5) / 2) ^ 102 - ((1 - Real.sqrt 5) / 2) ^ 102)
:= by sorry
```

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

Compiled, yet misaligned

general > FormalMATH Benchmarking Formal Mathematical Reasoning of LLM — MAY 9

Zhouliang Yu 1:36 PM

Excited to introduce FormalMATH: a large-scale formal math benchmark with 5,560 formally verified Lean 4 statements from Olympiad and UG-level problems.
Best model performance: just 16.46% — plenty of room for progress!
We open-sourced the complete datasets with the one-click evaluation code.

Github: https://github.com/Sphere-AI-Lab/FormalMATH-Bench
Huggingface: https://huggingface.co/datasets/SphereLab/FormalMATH-All
Website: https://spherelab.ai/FormalMATH/
LeaderBoard: https://spherelab.ai/FormalMATH/#toggleButton

🎉 8

Jason Rute 6:48 PM

I worry about having too many benchmarks which are all the same. What do you see this benchmark adding over say PutnamBench?

The most interesting part of this benchmark is that it is IIUC automatically generated with a human in the loop. But that probably means a lot of mistakes. See the recent discussions on this Zulip about the CombiBench benchmark and the ProverBench benchmark for all the different ways that one can easily make mistakes with such benchmarks. The first mistake I found in your is in problem `quantitative_reasoning_zh_blue_41`. I think that is the forth problem I looked at. You made a common mistake forgetting that subtraction of natural numbers is truncated subtraction.

The problem is false for x = 2.

FormalMATH record: quantitative_reasoning_zh_blue_41 compiler feedback: no errors

Given $M = \{x \mid x = a^2+1, a \in \mathbb{N}^*\}$, $N = \{x \mid x = b^2-4b+5, b \in \mathbb{N}^*\}$, determine the relationship between M and N. Answer: $M \subseteq N$, since $a^2+1 = (a+2)^2-4(a+2)+5$, and $1 \in N$ but $1 \notin M$.

```
theorem quantitative_reasoning_zh_blue_41 :
    {x : ℕ | ∃ a : ℕ, 0 < a ∧ x = a^2 + 1} ⊂
    {x : ℕ | ∃ b : ℕ, 0 < b ∧ x = b^2 - 4 * b + 5} := by sorry
```

FormalMATH record: theorem_proving_zh_blue_41 compiler feedback: no errors

Let S be a subset of {1, 2, …, 50} such that the sum of any two distinct elements in S is not divisible by 7. What is the maximum possible number of elements in S?
Reference solution: $|S| = |K_1|+|K_2|+|K_3|+1 = \mathbf{23}$.

```
def property (s : Finset ℕ) : Prop := ∀ x ∈ s, ∀ y ∈ s, x ≠ y → ¬(7 ∣ x + y)

theorem theorem_proving_zh_blue_41 :
    IsGreatest {n | ∃ s : Finset ℕ, s ⊆ Finset.Icc 1 50 ∧ property s ∧ n = s.card} 16 := by sorry
```

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

Compiled, yet misaligned

Semantic Verification via LLMs. We implement a semantic verification strategy based on multiple powerful general-purpose LLMs (*e.g.*, o1-mini [JKL+24], claude-3.5-Sonnet) to evaluate semantic alignment between natural language mathematics problems and their Lean4 formalizations. Each model employs chain-of-thought reasoning (See the prompt in Section G) to complete the following procedures: (1) back-translate Lean4 statements into natural language, (2) compare reconstructed descriptions with original problems, and (3) provide binary judgments (*i.e.*, aligned/misaligned). Importantly, only Lean4 statements that passed semantic verification performed by all the LLMs would be collected. This strategy is guided by the insight that translating Lean4 statements to natural language is a much easier task than the reverse process, and general-purpose LLMs excel at understanding natural language phrasings [WJL+22]. Overall, this procedure filters out 60.7% of syntactically correct but semantically misaligned statements (*i.e.*, from 92.4% to 32.7%). Interestingly, we find distinct consensus patterns across problem difficulty levels – around 30% unanimous agreement rate for high school competition problems and significantly lower consensus for undergraduate-level formalizations (*e.g.*, 4.63% on HardMath).

Disproving a Statement by Proving Its Negation. Inspired by the Law of the Excluded Middle (LEM [Wik25b]), we further filter out certain non-provable formalizations using off-the-shelf LLM-based provers (*e.g.*, DeepSeek-Prover-V1.5). For any formalized statement $\mathbf{T}_n^{(k)}$, we perform the following steps: (1) construct logical negation: construct its logical negation by applying transformation rules such as De Morgan dualization [Wik25a] to generate $\neg\mathbf{T}_n^{(k)}$, and (2) automated proof attempts: perform automated proof attempts on $\neg\mathbf{T}_n^{(k)}$ within the formal system $\mathcal{S}$ (*i.e.*, Lean4 compiler). A successful proof of $\neg\mathbf{T}_n^{(k)}$ implies that the original statement $\mathbf{T}_n^{(k)}$ cannot hold on $\mathcal{S}$. Example 3.2 illustrates the Lean 4 formalization of a number-theoretic conjecture and its negation. By constructing the negation of a statement and applying an LLM-based prover for disproof, the system identifies inconsistencies through boundary case testing (*e.g.*, n = 7) and derives contradictions via systematic case analysis (*i.e.*, `interval_cases`). This strategy has filtered out a few unprovable statements, accounting for 1.6% of the total statements.

Expert Verification. We have recruited 12 International Mathematical Olympiad medalist-level human experts to manually check the semantic alignment fidelity between natural language statements and their Lean4 formalizations. Table 2 shows some relevant information about the human validation stage. Our results show that our multi-LLM autoformalization and validation pipeline delivers substantial efficacy, retaining 72.1% of statements from the last stage of LLM-based semantic verification (from 30.1% to 21.7%) while significantly reducing manual verification efforts. In total, we have successfully formalized 21.7% of syntactically and semantically correct mathematical statements from a diverse collection of mathematical problems collected from multiple data sources. See Appendix A,C for more details.

Item	Value
# Annotators	12
Preservation rate	72.09%
Cost/statement	$6.89
Total duration	22 days

Table 2: Annotation statistics.

SphereLab M-A-P

FormalMATH: Benchmarking Formal Mathematical Reasoning of Large Language Models

Zhouliang Yu[1,2,*], Ruotian Peng[3,*], Keyi Ding[4,*], Yizhe Li[5], Zhongyuan Peng[4], Minghao Liu[5], Yifan Zhang[6], Zheng Yuan[4], Huajian Xin[4], Wenhao Huang[4], Yandong Wen[3], Ge Zhang[4], Weiyang Liu[7,†]

[1]The Chinese University of Hong Kong [2]Numina [3]Westlake University [4]M-A-P [5]2077AI [6]University of California, Los Angeles [7]Max Planck Institute for Intelligent Systems, Tübingen

FormalMATH record: quantitative_reasoning_zh_blue_41 compiler feedback: no errors

Given $M = \{x \mid x = a^2+1, a \in \mathbb{N}^*\}$, $N = \{x \mid x = b^2-4b+5, b \in \mathbb{N}^*\}$, determine the relationship between M and N. Answer: $M \subseteq N$, since $a^2+1 = (a+2)^2-4(a+2)+5$, and $1 \in N$ but $1 \notin M$.

```
theorem quantitative_reasoning_zh_blue_41 :
    {x : ℕ | ∃ a : ℕ, 0 < a ∧ x = a^2 + 1} ⊂
    {x : ℕ | ∃ b : ℕ, 0 < b ∧ x = b^2 - 4 * b + 5} := by sorry
```

FormalMATH record: theorem_proving_zh_blue_41 compiler feedback: no errors

Let S be a subset of {1, 2, ..., 50} such that the sum of any two distinct elements in S is not divisible by 7. What is the maximum possible number of elements in S? Reference solution: $|S| = |K_1|+|K_2|+|K_3|+1 = \mathbf{23}$.

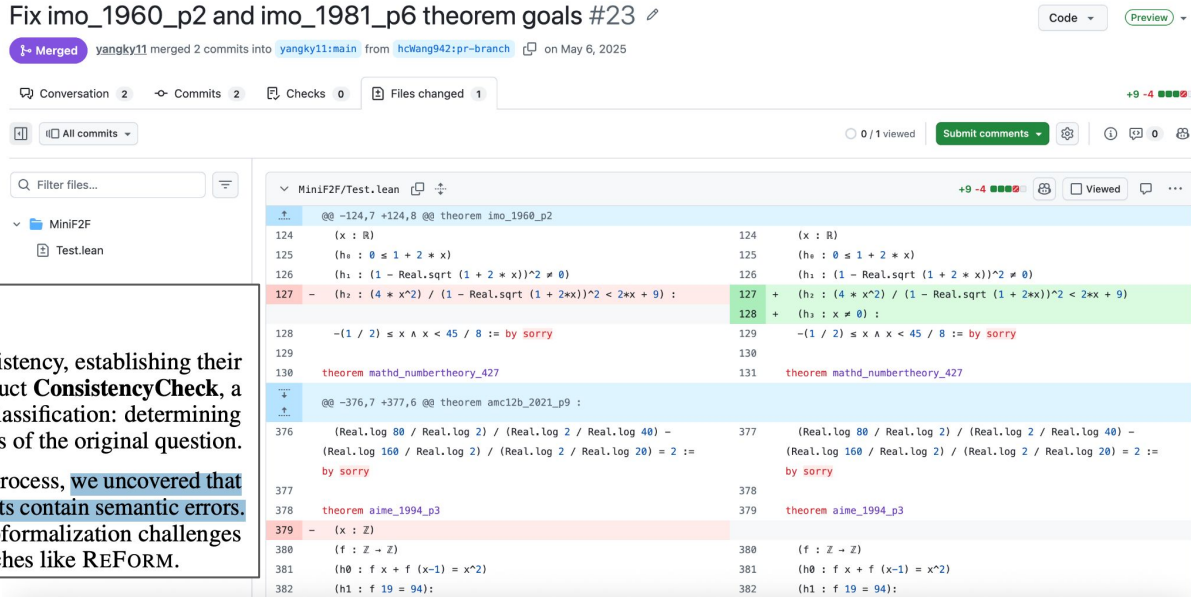
```
def property (s : Finset ℕ) : Prop := ∀ x ∈ s, ∀ y ∈ s, x ≠ y → ¬(7 ∣ x + y)

theorem theorem_proving_zh_blue_41 :
    IsGreatest {n | ∃ s : Finset ℕ, s ⊆ Finset.Icc 1 50 ∧ property s ∧ n = s.card} 16 := by sorry
```

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

Compiled, yet misaligned



4.5 Reliability of Semantic Consistency Evaluation

Since our evaluation relies on LLM-based judges to assess semantic consistency, establishing their reliability is crucial for validating our experimental conclusions. We construct **ConsistencyCheck**, a benchmark of 859 expert-annotated items where models perform binary classification: determining whether a formal statement correctly preserves the mathematical semantics of the original question.

Human expert fallibility in existing benchmarks. During the annotation process, we uncovered that 16.4% of miniF2F and 38.5% of ProofNet's human-written formal statements contain semantic errors. This high error rate in expert-crafted formalizations underscores that autoformalization challenges even human specialists, further motivating the need for automated approaches like REFORM.

- Expert-written benchmarks contain such errors: this PR fixes the theorem goals of two IMO problems in MiniF2F
- ReForm (2025): 16.4% of miniF2F and 38.5% of ProofNet human-written formal statements contain semantic errors

ReForm [Chen, et al., 2025]
https://arxiv.org/pdf/2510.24592

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

Compiled, yet misaligned

REFORM: Reflective Autoformalization with Prospective Bounded Sequence Optimization

Guoxin Chen[1,2,*], Jing Wu[2,*,†], Xinjie Chen[2,3], Wayne Xin Zhao[1,†], Ruihua Song[1,†], Chengxi Li[2], Kai Fan[2], Dayiheng Liu[2], Minpeng Liao[2,†]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Tongyi Lab, Alibaba Group, [3]Zhejiang University
{gx.chen.chn, batmanfly}@gmail.com, songruihua_bloon@outlook.com
{lz.wujing, chenxinjie.cxj, minpeng.lmp}@alibaba-inc.com
Model Benchmark Code

[Chen, et al., 2025]
https://arxiv.org/pdf/2510.24592

Data Format

Each record has the following JSON structure:

```
{
  "name": "problem_identifier",
  "split": "valid|test",
  "goal": "Lean4 goal statement",
  "header": "Lean4 imports and opening commands",
  "informal_statement": "Natural language problem statement",
  "formal_statement": "Formalized theorem statement",
  "human_check": "true|false",
  "human_reason": "Explanation for incorrect labels"
}
```

⚠ Known Issues

During annotation, we identified several problematic informal statements:

miniF2F Issues:

- `amc12a_2011_p18`: Missing specification of whether x equals zero
- `amc12_2000_p11`: Contains only answer choices without actual problem statement

ProofNet Issues:

- `exercise_1998_a3`: Incomplete condition after "such that"
- `exercise_1_18b`: Missing specification of whether x equals zero

informal_statement string · lengths 183↔255 14.4%	formal_statement string · lengths 343↔402 0.9%	human_check string · classes false 33.6%	human_reason string · lengths 144↔152 0.1%
Prove that if f is holomorphic in the unit disc, bounded and not identically zero, and $z_{1}, z_{2}, \ldots, z_{n}, \ldots$ are its zeros $\left(\left\|z_{k}\right\|<1\right)$, then $\sum_{n}\left(1-\left\|z_{n}\right\|\right)<\infty$.	theorem exercise_5_1 (f : ℂ → ℂ) (hf : DifferentiableOn ℂ f (Metric.ball 0 1)) (hb : Bornology.IsBounded (Set.range f)) (h0 : f ≠ 0) (zeros : ℕ → ℂ) (hz : ∀ n, f (zeros n) = 0) (hzz : Set.range zeros = {z | f z = 0 ∧ z ∈ (Metric.ball (0 : ℂ) 1)}) : ∃ (z : ℂ), Filter.Tendsto (λ n => (∑ i ∈ Finset.range n, (1 - zeros i))) atTop (𝓝 z) :=	false	hb is wrong, f is only bounded in the unit disc. h0 is wrong f is not identically zero in the unit disc. Goal is wrong, `zeros i` should be `‖zeros i‖`.
Prove that the ring $\mathbb{Z}\left[x_{1}, x_{2}, x_{3…	theorem exercise_9_1_10 {f : ℕ → MvPolynomial ℕ ℤ} (hf :…	false	`MvPolynomial.X i` in hf should be…
Let f be a real function on the real line with continuous third…	theorem exercise_1998_a3 (f : ℝ → ℝ) (hf : ContDiff ℝ 3 f…	false	This is an improper informal statement…
$ \textbf{(A)}\ 5 \qquad \textbf{(B)}\ 6 \qquad \textbf{(C)}…	theorem amc12a_2011_p18 (x y : ℝ) (h₀ : abs (x + y) + ab…	false	This is an improper informal statement…
If $k = 1$ and $\mathbf{x} \in R^{k}$, prove that there does not…	theorem exercise_1_18b : ¬ ∀ (x : ℝ), ∃ (y : ℝ), y ≠ 0 ∧…	false	This is an improper informal statement…

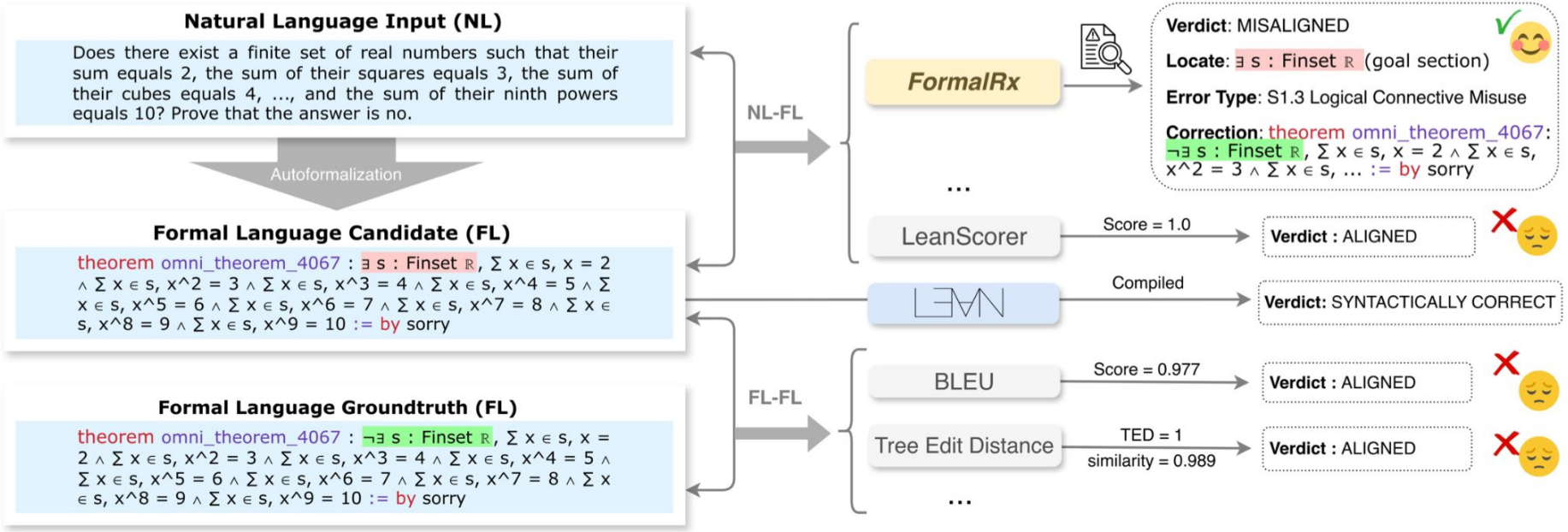
Related Works

Table 1. Comparison of semantic evaluation methods for autoformalization. We categorize the diagnosis components as follows: ***Verdict***: the alignment decision on the autoformalization output; ***Categorize***: a fine-grained classification of the error type based on the proposed SCI error taxonomy; ***Locate***: the error location (if misaligned); ***Rectify***: the corrected formal statement output (if misaligned).

Compa-rison	Source Paper	Type	Method	Compositional Diagnosis			
				Verdict	Categorize	Locate	Rectify
FL-FL	ProofNet (Azerbayev et al., 2023)	Naive Matching	BLEU	Binary	✗	✗	✗
	ProofNet (Azerbayev et al., 2023)	Naive Matching	Typecheck	Binary	✗	✗	✗
	BEq (Liu et al., 2025b)	Rule-based	Equivalence Metric	Binary	✗	✗	✗
	ProofBridge (Jana et al., 2025)	LLM+Rule-based	Equivalence Metric	Binary	✗	✗	✗
	GTED (Liu et al., 2025d)	Rule-based	Tree Edit Distance	Binary	✗	✗	✗
	TransTED (Liu et al., 2025c)	Rule-based	Tree Edit Distance	Binary	✗	✗	✗
	LeanEuclid (Murphy et al., 2024)	Rule-based	SMT Solver	Binary	✗	✗	✗
NL-NL	LeanWorkbook (Ying et al., 2024)	Rule-based	NLI Test	Binary	✗	✗	✗
	Herald (Gao et al., 2025)	Rule-based	NLI Test	Binary	✗	✗	✗
	FormalMATH (Yu et al., 2025b)	LLM-as-Judge	Multi-LLM Ensemble Voting	Binary	✗	✗	✗
NL-FL	Lean Workbook (Ying et al., 2024)	LLM-as-Judge	Multi-LLM Majority Voting	Binary	✗	✗	✗
	ATF (Guo et al., 2025b)	LLM-as-Judge	Multi-LLM Ensemble Voting	Binary	✗	✗	✗
	ReForm (Chen et al., 2025)	LLM-as-Judge	Single-LLM	Binary	✗	✗	✗
	LeanScorer (Yu et al., 2025a)	LLM-as-Judge	Subtask Decomposition	Binary	✗	✗	✗
	AriaScorer (Wang et al., 2025a)	LLM-as-Judge	Subtask Decomposition	Binary	✗	✗	✗
	FormalAlign (Lu et al., 2025)	Training-based	SFT (CE+Contrastive Loss)	Binary	✗	✗	✗
	FORMALRX (Ours)	Training-based	SFT (CE Loss)	Multiple	✓	✓	✓

FormalRx return a Diagonosis

😢 **Lack of automated detection & correction tool.**



FormalRx replaces that single verdict with 4 actionable diagnosis in one forward pass.

Task Definition

Given an informal statement S and its candidate formalization F, FormalRx treats evaluation as conditional generation: it maps (S, F) to a structured diagnostic sequence Y with four components:

Component	Description
Verdict	Binary judgment (**Aligned** / **Misaligned**)
Categorization	Specific error type from the SCI Error Taxonomy
Localization	Problematic code segment within F
Correction	Rectified code F' semantically aligned with S

SCI Error Taxonomy

FormalRx classifies every autoformalization error into *SCI Error Taxonomy*, a hierarchical scheme that contains 28 categories along three dimensions.

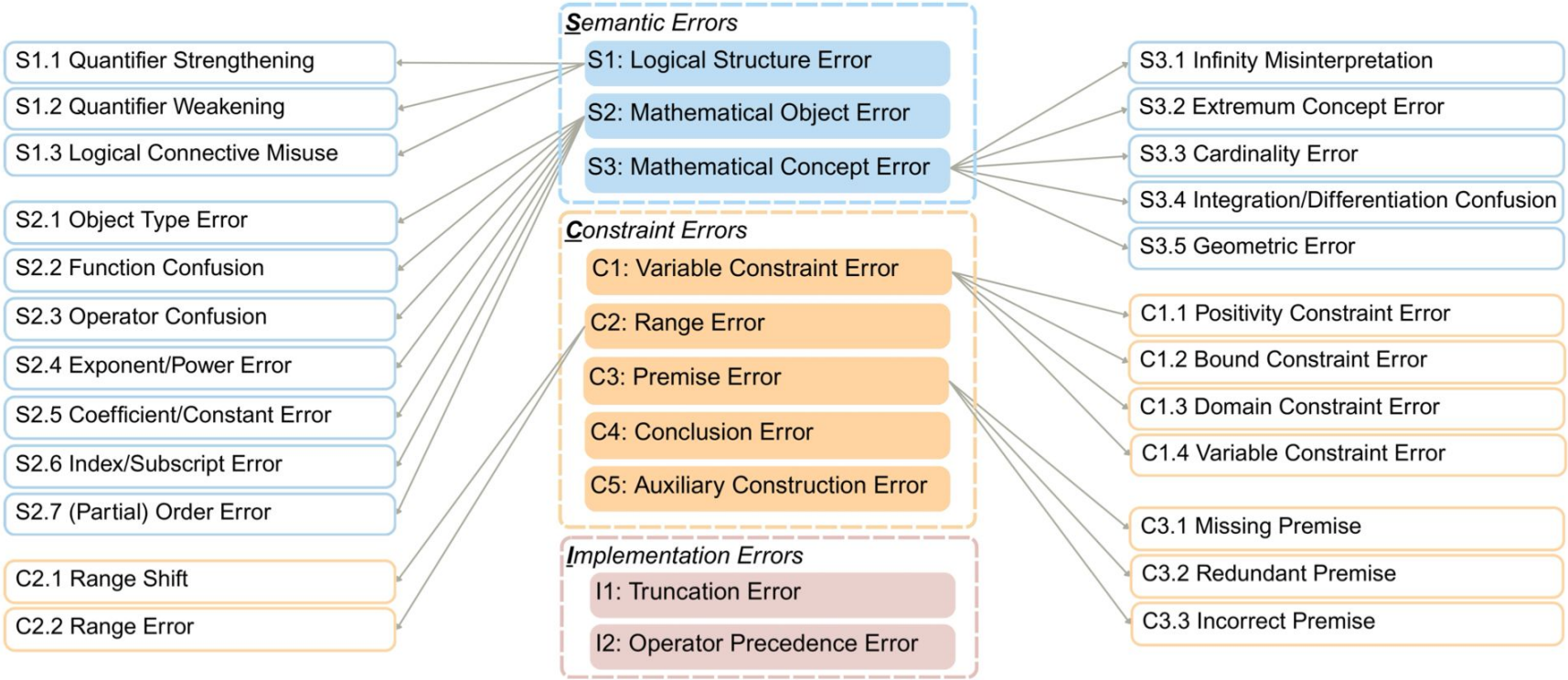
- *S*emantic Errors in mathematical content,
- *C*onstraint Errors in the conditions under which a statement holds, and
- *I*mplementation Errors in Lean 4 execution semantics.

Partition-based Design. Categories are pairwise disjoint and collectively exhaustive, guarantees that every error admits a deterministic single-label assignment.

Pirority Ordering. Specific error types take precedence over the location-based catch-alls (premise, conclusion, auxiliary), and smaller indices win within each category, ensuring mutual exclusivity.

Iterative Validation. Taxonomy was developed through iterative cycles of pattern identification, category abstraction, and empirical validation.

SCI Error Taxonomy



香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

FormalRx return a Diagonosis

theorem_proving_zh_blue_41 compiles: no errors

Let $S \subseteq \{1, \ldots, 50\}$ such that the sum of any two distinct elements of S is not divisible by 7. Find the maximum $|S|$. Reference solution: **23**.

```
def property (s : Finset ℕ) : Prop :=
  ∀ x ∈ s, ∀ y ∈ s, x ≠ y → ¬(7 ∣ x + y)

theorem theorem_proving_zh_blue_41 :
    IsGreatest {n | ∃ s : Finset ℕ,
      s ⊆ Finset.Icc 1 50 ∧ property s
      ∧ n = s.card} 16 := by sorry
```

↓ FormalRx ↓

Verdict: MISALIGNED

Error Type: S2.5 Coefficient/Constant Error

Location: `16` (goal constant)

Correction:

```
def property (s : Finset ℕ) : Prop :=
  ∀ x ∈ s, ∀ y ∈ s, x ≠ y → ¬(7 ∣ x + y)

theorem theorem_proving_zh_blue_41 :
    IsGreatest {n | ∃ s : Finset ℕ,
      s ⊆ Finset.Icc 1 50 ∧ property s
      ∧ n = s.card} 23 := by sorry
```

quantitative_reasoning_zh_blue_41 compiles: no errors

Given $M = \{x \mid x = a^2+1, a \in \mathbb{N}^*\}$, $N = \{x \mid x = b^2-4b+5, b \in \mathbb{N}^*\}$, determine their relationship. Answer: $M \subsetneq N$.

```
theorem quantitative_reasoning_zh_blue_41 :
    {x : ℕ | ∃ a : ℕ, 0 < a ∧ x = a^2 + 1}
    ⊂ {x : ℕ | ∃ b : ℕ, 0 < b ∧
       x = b^2 - 4 * b + 5} := by sorry
```

↓ FormalRx ↓

Verdict: MISALIGNED

Error Type: I1 Truncation Error

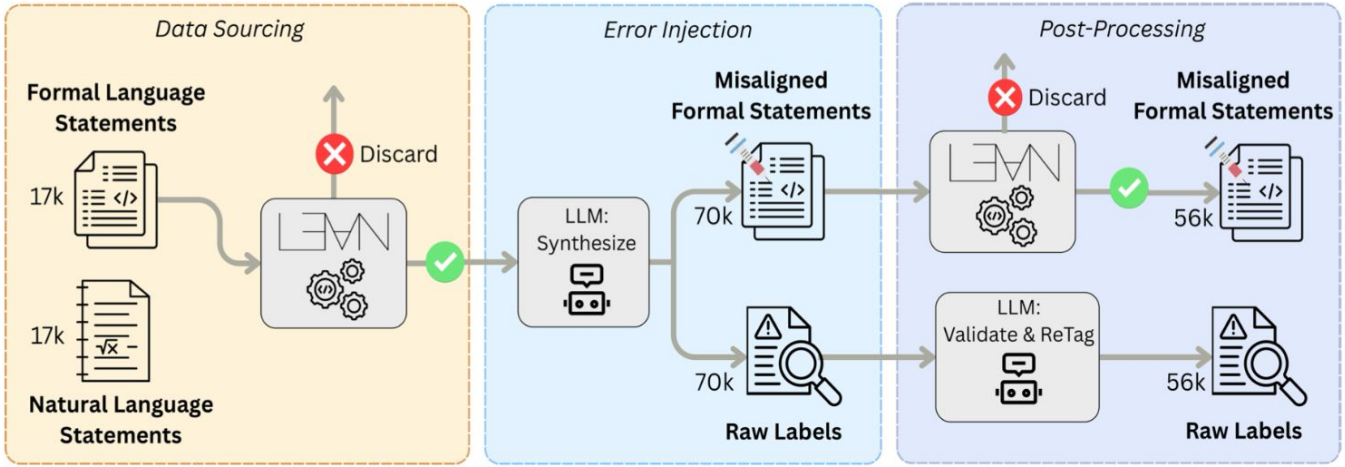
Location: `b^2 - 4 * b + 5` (over ℕ)

Correction:

```
theorem quantitative_reasoning_zh_blue_41 :
    {x : ℤ | ∃ a : ℤ, 0 < a ∧ x = a^2 + 1}
    ⊂ {x : ℤ | ∃ b : ℤ, 0 < b ∧
       x = b^2 - 4 * b + 5} := by sorry
```

Data Synthesis

Category	Source	Count
Dataset	FIMO	149
	Compfiles	245
Benchmark	PutnamBench	659
	ProverBench	325
	CombiBench	101
	Gaokao-Formal	495
	Formal-MATH	5,505
Library	Mathlib	10,346
Total		**17,825**



7,479
from datasets & benchmarks

10,346
doc-string pairs from Mathlib

100%
compile under Lean v4.24.0

0
from MiniF2F / ProofNet

Data Synthesis

Putnam 2013 · B1

Seed Pair (aligned)

For positive integers n, define $c(1) = 1$, $c(2n) = c(n)$, and $c(2n+1) = (-1)^n c(n)$. Find $\sum_{n=1}^{2013} c(n)\,c(n+2)$.

```
theorem putnam_2013_b1 (c : ℕ → ℤ) (hc1 : c 1 = 1)
    (hceven : ∀ n : ℕ, n > 0 → c (2 * n) = c n)
    (hcodd : ∀ n : ℕ, n > 0 → c (2 * n + 1) = (-1) ^ n
    * c n) :
    (∑ n : Set.Icc 1 2013, c n * c (n.1 + 2)) = (-1 :
    ℤ) := by sorry
```

LLM: Synthesize, inject C2.1 Range Shift

Misaligned Formal Statement

```
theorem putnam_2013_b1 (c : ℕ → ℤ) (hc1 : c 1 = 1)
    (hceven : ∀ n : ℕ, n > 0 → c (2 * n) = c n)
    (hcodd : ∀ n : ℕ, n > 0 → c (2 * n + 1) = (-1) ^
    n * c n) :
    (∑ n : Set.Icc 0 2013, c n * c (n.1 + 2)) = (-1 :
    ℤ) := by sorry
```

Category: C2.1 Range Shift

Location: `Set.Icc 0 2013`

Reasoning: *The hypotheses pin down c only for $n \geq 1$, so the stray $n = 0$ term $c(0)\,c(2)$ is completely unconstrained: the sum need not equal -1 for every admissible c, and the statement becomes unprovable.*

assemble

Putnam 2013 · B1

Natural Language

For positive integers n, define $c(1) = 1$, $c(2n) = c(n)$, and $c(2n+1) = (-1)^n c(n)$. Find $\sum_{n=1}^{2013} c(n)\,c(n+2)$.

Formal Language

```
theorem putnam_2013_b1 (c : ℕ → ℤ) (hc1 : c 1 = 1)
    (hceven : ∀ n : ℕ, n > 0 → c (2 * n) = c n)
    (hcodd : ∀ n : ℕ, n > 0 → c (2 * n + 1) = (-1) ^ n
    * c n) :
    (∑ n : Set.Icc 0 2013, c n * c (n.1 + 2)) = (-1 :
    ℤ) := by sorry
```

Verdict: Misaligned ✗

Categorization: C2.1 Range Shift

Localization: `Set.Icc 0 2013`

Correction: ✓

```
theorem putnam_2013_b1 (c : ℕ → ℤ) (hc1 : c 1 = 1)
    (hceven : ∀ n : ℕ, n > 0 → c (2 * n) = c n)
    (hcodd : ∀ n : ℕ, n > 0 → c (2 * n + 1) = (-1) ^
    n * c n) :
    (∑ n : Set.Icc 1 2013, c n * c (n.1 + 2)) = (-1 :
    ℤ) := by sorry
```

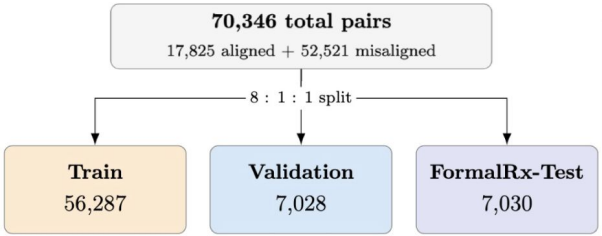
香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

https://huggingface.co/datasets/LARK-Lab/FormalRx-Test

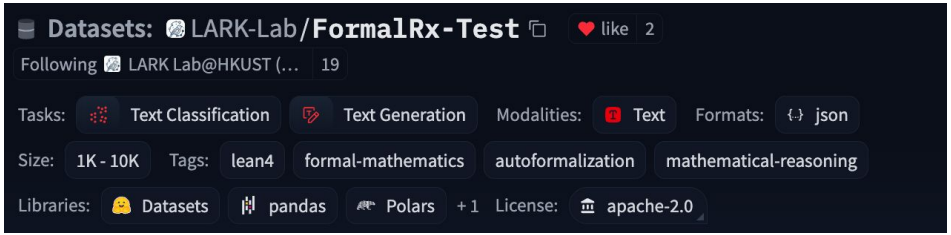
FormalRx Test

The held-out test set ***FormalRx-Test*** 🤗 is the first fine-grained diagnostic benchmark for autoformalization.



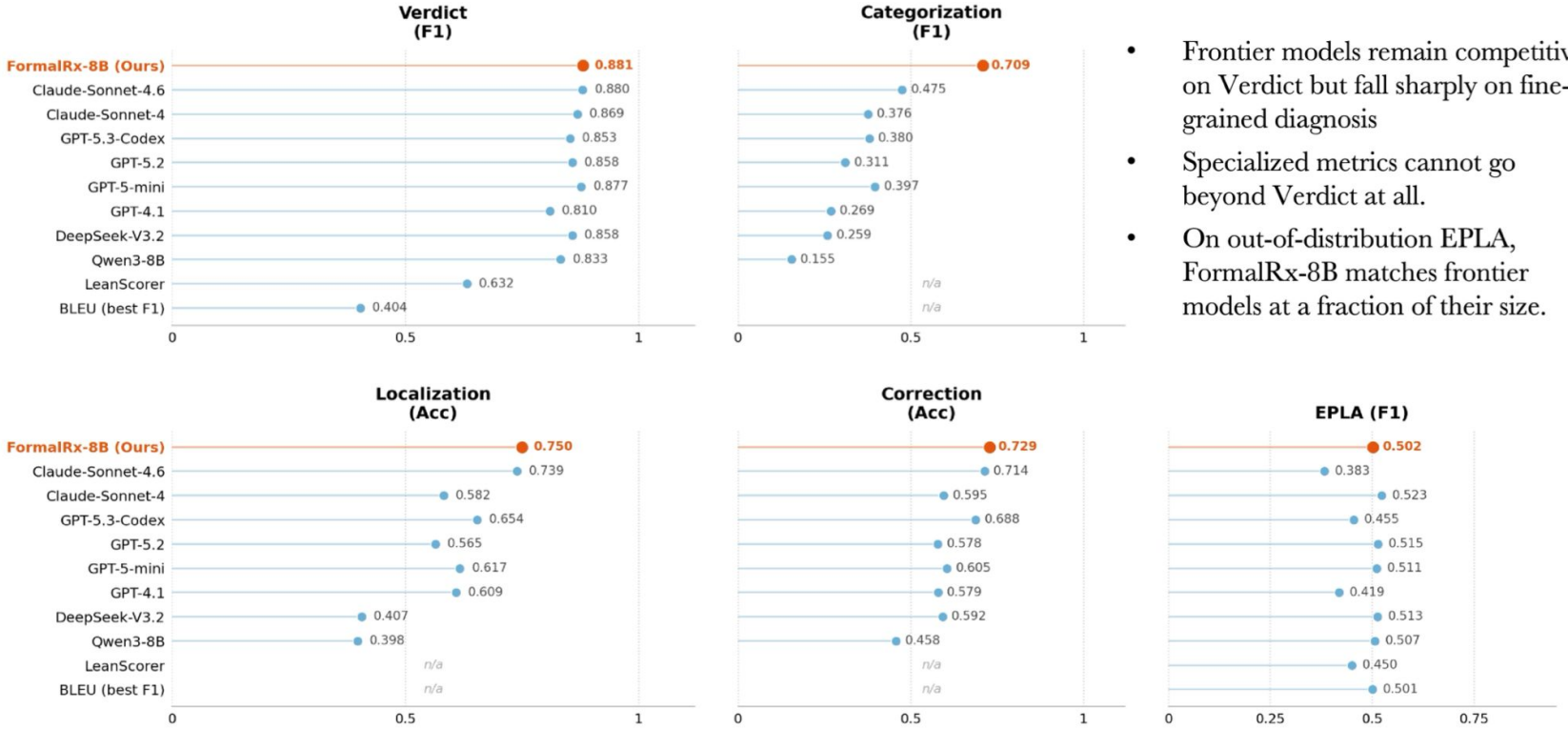
7,030	1,791	5,239	28
Total samples	Aligned (25.5%)	Misaligned (74.5%)	SCI categories covered

Since release, it has been downloaded 2,000+ times, and serves as the Challenge Track 1 benchmark at the 3rd AI4Math Workshop @ ICML2026, and has drawn wide attention from the community.



🔍Scan for Project

Result

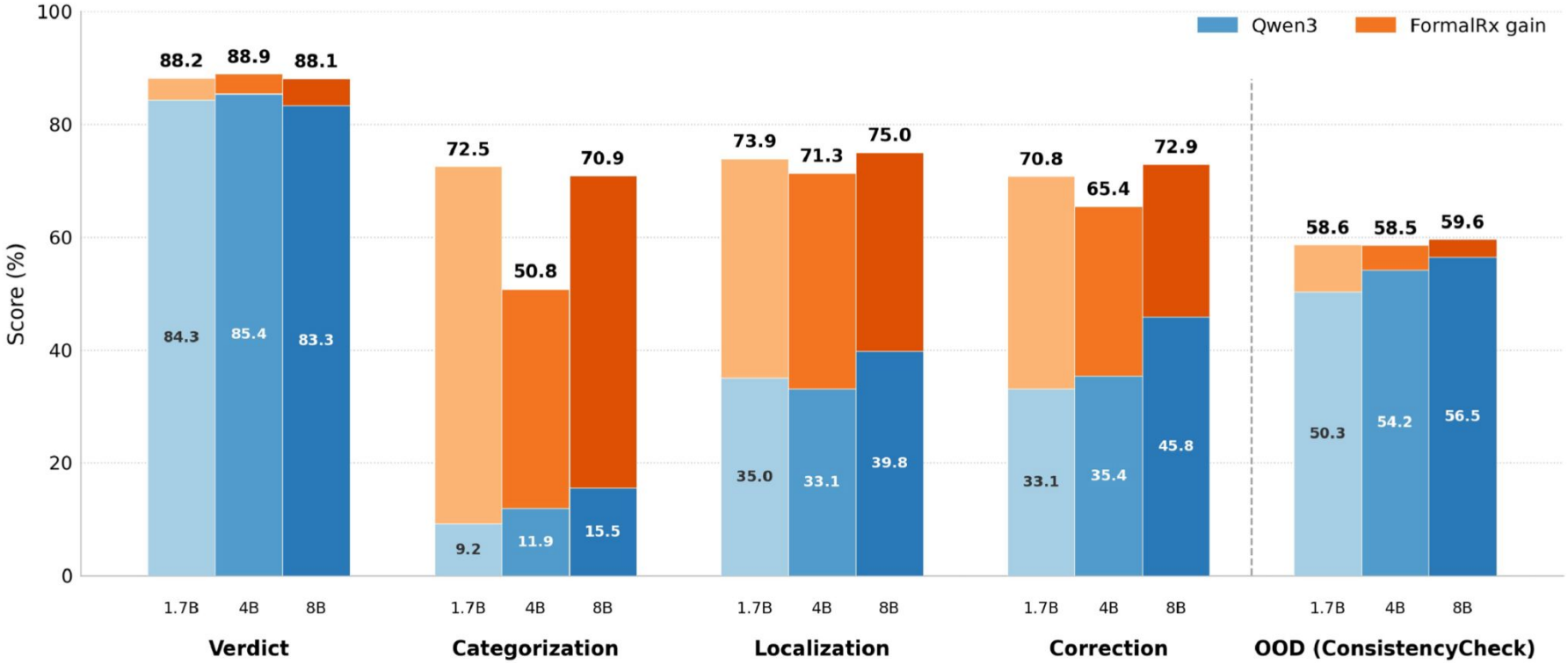


- Frontier models remain competitive on Verdict but fall sharply on fine-grained diagnosis
- Specialized metrics cannot go beyond Verdict at all.
- On out-of-distribution EPLA, FormalRx-8B matches frontier models at a fraction of their size.

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

Scaling Analysis



Downstream Application

Metric	R0	FormalRx R1	FormalRx R2	FormalRx R3	LS R1	LS R2	LS R3
Compile Rate (by cand)	51.4%	95.7%	96.7%	98.5%	95.8%	97.3%	97.2%
Compile Rate (by prob)	75.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Pass@4 (loc.)	35.0%	44.7%	42.2%	39.5%	30.3%	26.1%	29.6%
Pass@4 (acc.)	35.0%	40.0%	41.0%	41.0%	36.0%	36.0%	37.0%
Verifier Prec. (by cand)	46.3%	26.3%	36.9%	–	32.9%	29.0%	–
Verifier Prec. (by prob)	60.0%	37.9%	44.0%	–	31.2%	35.0%	–
Problems refined	47	45	43	–	23	23	–
Gains vs R0	0	4	6	3	0	1	1
Losses vs R0	0	2	2	2	5	2	3
Net vs R0	+0	+2	+4	+1	−5	−1	−2

Limitations

- Training distribution is competition and Mathlib style; research-level statements are the next stress test
- Out-of-domain evaluation is verdict-only today; no fine-grained OOD benchmark exists yet
- Self-refinement evidence is n = 100; the mechanism is controlled, scaling it up is future work

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

LAanguage Reasoning & Knowledge Lab

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

ETH zürich

UCLA

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

ETH AI CENTER

Key Takeaways

- Compiled does not mean aligned
- Evaluation as diagnosis
  - Binary verdict cannot close that gap
  - Unified feedback over SCI Taxonomy
  - Minimal component on error location
  - Trainable at 1.7B/4B/8B, beating frontier models where explanation matters
- Diagnosis signal with richer feedback improves refinement
  - Unified and dense reward for RL is what we build next

FORMALRX: Rectify and eXamine Semantic Failures in Autoformalization

Haocheng Wang [*1,2] Baiyu Huang [*1] Yingjia Wan [*3] Xiao Zhu [1] Xiaoyang Liu [4]
Yinya Huang [†5] Zhijiang Guo [†1,6]

Scan for Project

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

*LA*anguage *R*easoning & *K*nowledge Lab

Thanks for your attention!

🌐 https://haocheng.wang/

📮 hcwang942@gmail.com

🔍Scan for Project