Lean4PHYS: Comprehensive Reasoning Framework for College-level Physics in Lean4

Yuxin Li*, Minghao Liu*, Ruida Wang*, Wenzhao Ji,
Zhitao He, Rui Pan, Junming Huang, Tong Zhang, Yi R. Fung†

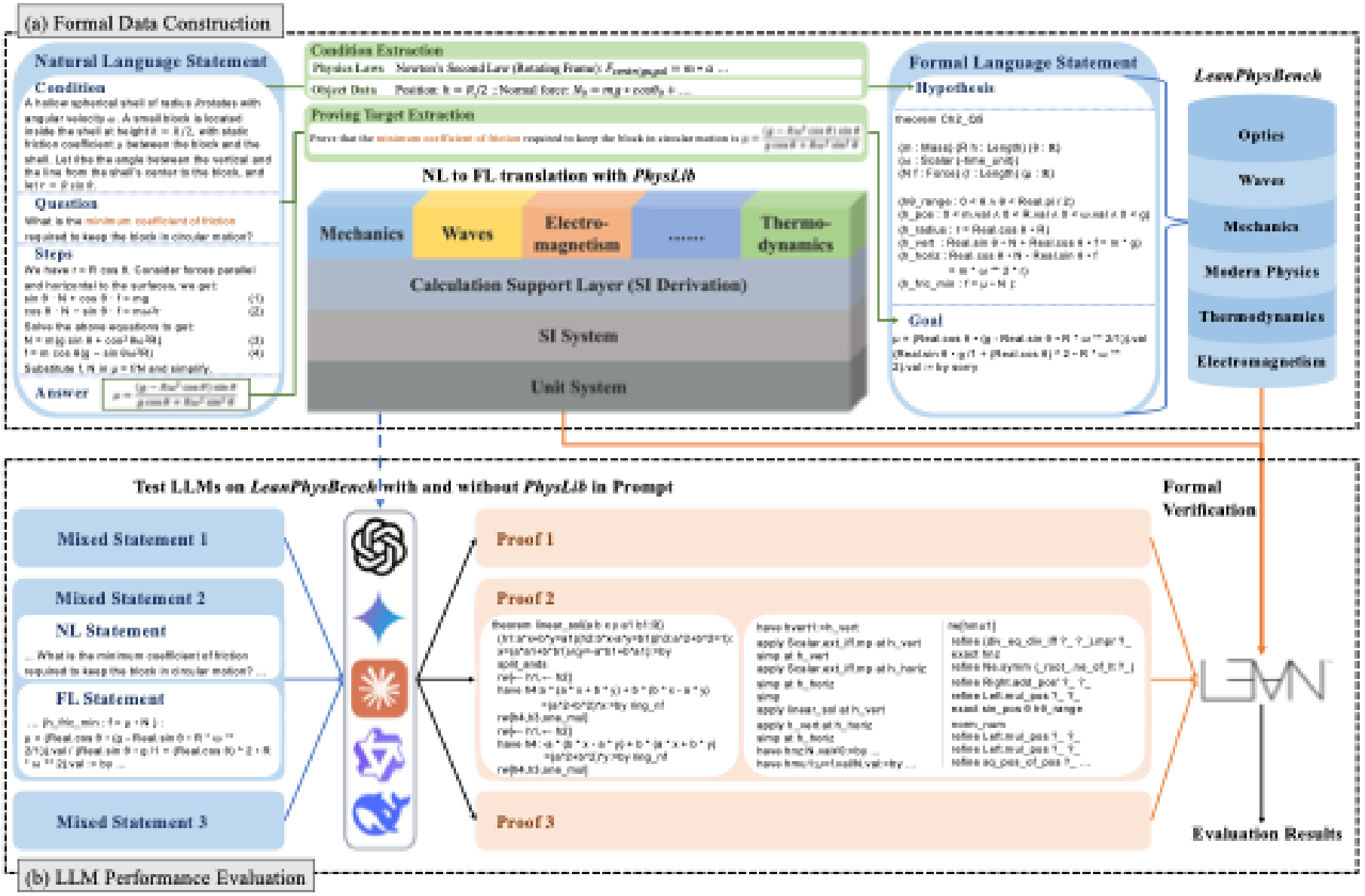
ICLR 2026

Lean4PHYS Background Introduction

- **Formal reasoning** is a key objective in machine learning often evaluated through rigorous mathematical derivations.
- **Formal Languages (FLs)** enable the automatic verification of both the reasoning process and its results.

- **Superior Results on Lean4 math benchmarks:**
  - Whether such formal capability can be transferred to other domains?
  - Or current provers are overfitted on mathematical reasoning.

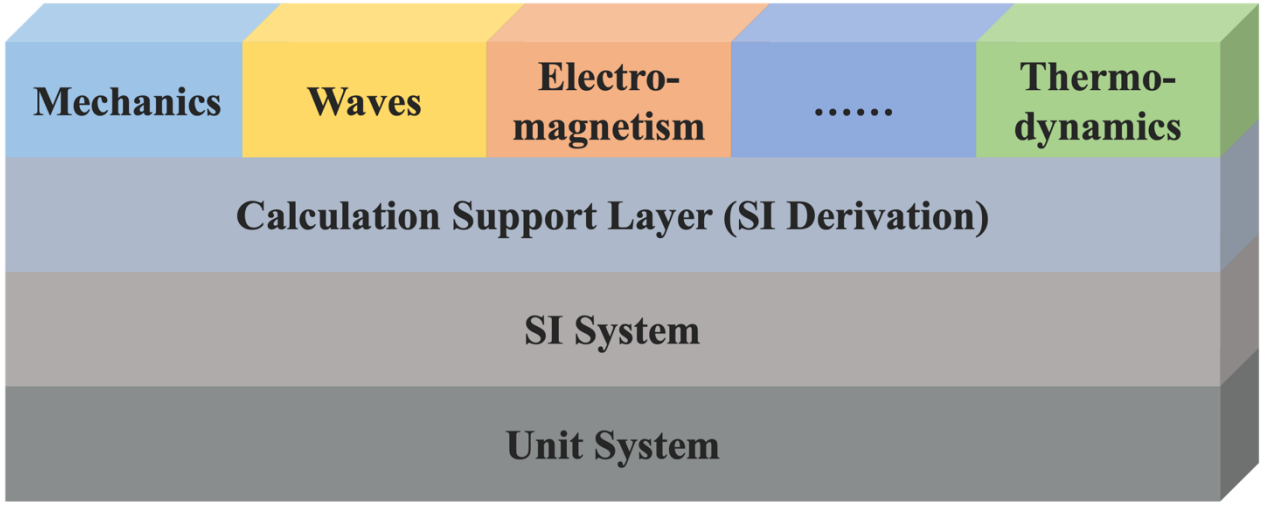
Lean4PHYS Overview



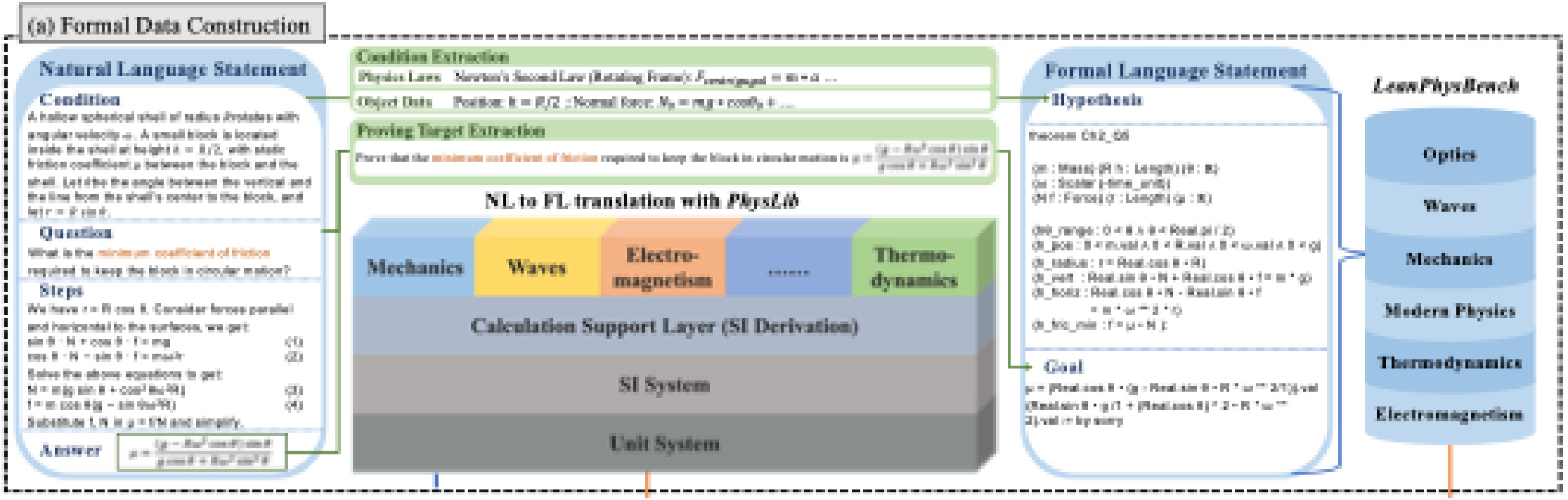
Formal Data Construction: *PhysLib*

- How to Formalize Physics Problems?

- The structure of *PhysLib*:
  - The Foundation System
  - Topic-Based Theorem System



Formal Data Construction



- ***PhysLib*** **-** a community-driven repository that includes fundamental unit systems and theorems essential for formal physics reasoning
- ***LeanPhysBench*** **-** a college-level benchmark for Lean4 formal physics reasoning

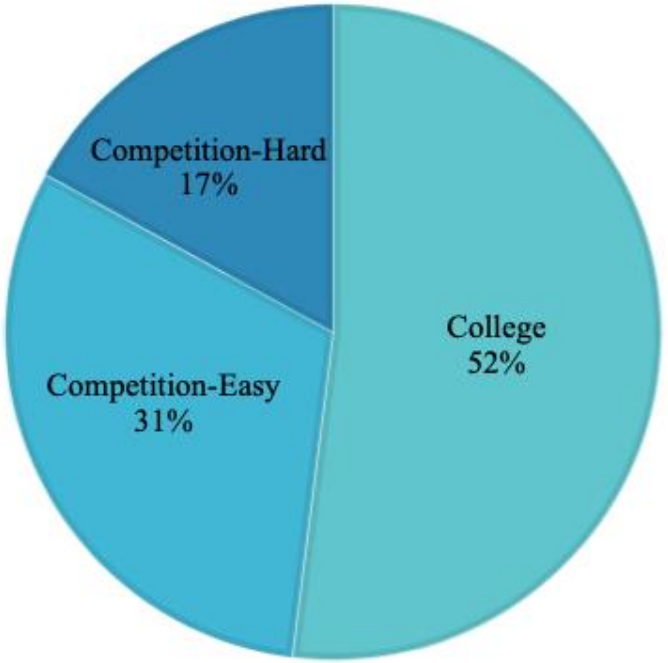
Formal Data Construction: *LeanPhysBench*

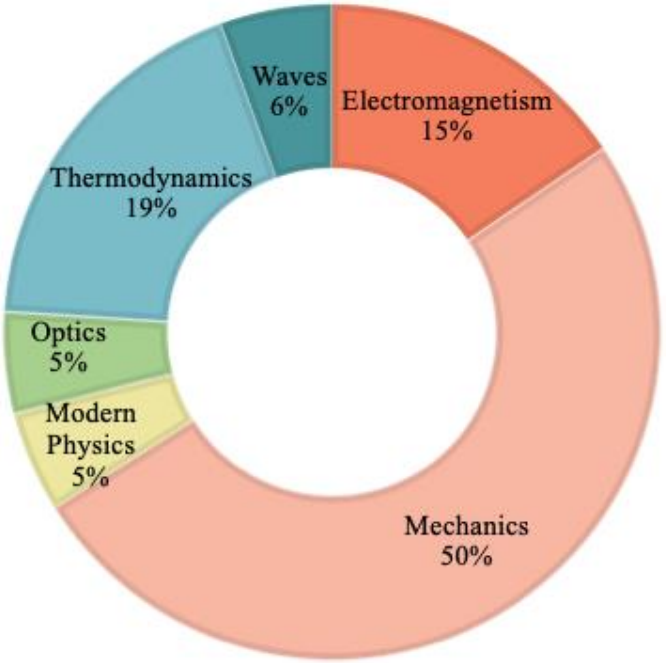
University Textbook Level

```
theorem University_Mechanics_3
  (x_0 x_1 : Length)
  (t_0 t_1 dt t : Time)
  (v_0 v_1 : Speed)
  (a: Acceleration)
  (xf xf1: Time → Length)
  (vf : Time → Speed)
  (ht0 : t_0 = 0 · second)
  (ht1 : t_1 = 4 · second)
  (ht : dt = t_1 - t_0)
  (ha : a = (v_1 -v_0)/dt)
  (hv : ∀ t, vf t = v_0 + a * t/1)
  (hv1: v_1 = vf dt)
  (hx: ∀ t, xf t = (a * t**2)/2 + v_0 * t)
  (hxx: ∀ t, xf1 t = (3 · meter / second**2)* t**2 - 2 · meter /
    second * t)
  (hxxx: xf = xf1):
  (a = 6 · meter / second**2 ∧ v_0 = -2 · meter / second) := by sorry
```

Olympiad Competition Level

```
theorem competition_mechanics_Ch2_Q32
  (m : Mass)
  (R : Length)
  (θ : ℝ)
  (v : Speed)
  (μ : ℝ)
  (N f : Force)
  (h_pos : 0 < m.val ∧ 0 < μ ∧ 0 < g.val)
  (h_sin_cos : Real.sin θ ≠ 0 ∧ Real.cos θ ≠ 0)
  (r_def : r = Real.sin θ · R)
  (h_horiz : Real.sin θ · N - Real.cos θ · f = m * v**2 / r)
  (h_vert  : Real.cos θ · N + Real.sin θ · f = m * g)
  (f_def : f = m * (Real.sin θ · g - Real.cos θ · v**2 / r / 1))
  (N_def : N = m * (Real.cos θ · g + Real.sin θ · v**2 / r / 1))
  (fric_bound : ‖f.val‖ ≤ μ * ‖N.val‖) :
  ∀ (ε : ℤ), (ε = 1 ∨ ε = -1) →(f = (ε : ℝ) · μ · N → v**2 =
    ((Real.sin θ - (ε : ℝ) * μ * Real.cos θ) · g * (Real.sin θ · R)\
    (Real.cos θ + (ε : ℝ) * μ * Real.sin θ))) := by sorry
```





- **Textbook-Level:** simple calculations, basic estimations, direct formula use.
- **Competition-Level:** derivations, advanced math, multi-step reasoning, deeper understanding.

LLM Performance Evaluation

Method	PhysLib	College	Comp-Easy	Comp-Hard	Overall
Open-source models					
DeepSeek-R1-8B	✓	6.09% ± 0.45%	8.60% ± 0.76%	0.00% ± 0.00%	5.83% ± 0.47%
	✗	2.56% ± 1.81%	5.91% ± 0.76%	0.98% ± 1.39%	3.33% ± 0.94%
Qwen3-8B	✓	6.41% ± 0.91%	10.22% ± 1.52%	0.00% ± 0.00%	6.52% ± 0.81%
	✗	2.58% ± 0.47%	3.23% ± 0.00%	0.00% ± 0.00%	2.35% ± 0.25%
Kimina-Prover-8B	✓	9.94% ± 0.91%	23.12% ± 1.52%	**6.86%** ± 1.39%	13.50% ± 0.71%
	✗	5.77% ± 0.00%	17.20% ± 0.76%	5.88% ± 0.00%	9.33% ± 0.24%
Goedel-Prover-V2-8B	✓	10.58% ± 1.36%	26.34% ± 1.52%	5.88% ± 0.00%	14.67% ± 1.18%
	✗	6.09% ± 0.45%	19.35% ± 0.00%	4.90% ± 1.39%	10.00% ± 0.00%
DeepSeek-Prover-V2-7B	✓	8.01% ± 0.45%	31.18% ± 1.52%	5.88% ± 0.00%	14.83% ± 0.24%
	✗	6.09% ± 0.45%	22.58% ± 0.00%	5.88% ± 0.00%	11.17% ± 0.24%
Closed-source models					
GPT-4o	✓	9.29% ± 2.40%	30.11% ± 2.01%	0.00% ± 0.00%	14.17% ± 0.85%
	✗	2.24% ± 0.45%	4.30% ± 2.01%	0.00% ± 0.00%	2.50% ± 0.41%
Claude-Sonnet-4	✓	29.49% ± 0.45%	61.83% ± 1.52%	0.00% ± 0.00%	34.50% ± 0.41%
	✗	3.21% ± 1.63%	5.91% ± 1.52%	0.00% ± 0.00%	3.50% ± 1.22%
Gemini-2.5-pro	✓	**32.69%** ± 0.79%	**74.19%** ± 1.32%	0.00% ± 0.00%	**40.00%** ± 0.71%
	✗	6.09% ± 0.45%	11.29% ± 1.32%	0.00% ± 0.00%	6.67% ± 0.71%

- Integrating ***PhysLib*** leads to **consistent performance improvements**.
- Found that **current expert math provers and general models** perform suboptimally, regardless of size. Observed potential **overfitting** to the math dataset for expert Lean provers.
- On the **hard dataset**, all models perform properly, highlighting **limitations in formal physics**, especially for statements involving **complex operations** like integrals and derivatives.

The End