PutnamBench: A Multilingual Competition-Mathematics Benchmark for Theorem Proving

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin,
Michelle Ding, Michael Jennings, Amitayush Thakur, Swarat Chaudhuri
UT Austin

ICML AI4Math Workshop: Best Paper Award

Motivation

- Want to benchmark olympiad-level mathematical reasoning
- MiniF2F (488)
  - Some problems from MATH, formalized
  - Some AMC/AIME/IMO problems
- FIMO (149)
  - IMO shortlist problems

Motivation II

- William Lowell Putnam Mathematical Competition:
  - taken by 1000s of **undergraduate** students yearly in North America
  - problems require knowledge from a **broad range of topics** in undergrad curriculum (analysis, abstract algebra, ..)
  - **Correlated with IMO**: IMO medalists are usually top performers

PutnamBench

- Formalizations of Putnam problems from competitions 1962 - 2023
- 640 formalized in Lean 4 & Isabelle, 417 formalized in Coq
- Many problems rely on mathematical theory developed in Mathlib, the HOL library, and various Coq repositories

PutnamBench

Benchmark	#	Natural Language	Lean	Isabelle	Coq	Factored Solution
miniF2F	488	✓	✓[†]	✓		
ProofNet	371	✓	✓[†]			N/A
Fimo	149	✓	✓[†]			
PutnamBench	640	✓	✓	✓	✓	✓

Putnam 2006 B2

Putnam 2006 B2. Prove that, for every set $X = \{x_1, x_2, \ldots, x_n\}$ of n real numbers, there exists a nonempty subset S of X and an integer m such that

$$\left| m + \sum_{s \in S} s \right| \leq \frac{1}{n+1}.$$

Putnam 2006 B2

Putnam 2006 B2. Prove that, for every set $X = \{x_1, x_2, \ldots, x_n\}$ of n real numbers, there exists a nonempty subset S of X and an integer m such that

$$\left| m + \sum_{s \in S} s \right| \leq \frac{1}{n+1}.$$

```
(a) theorem putnam_2006_b2
(n : ℕ)
(npos : n > 0)
(X : Finset ℝ)
(hXcard : X.card = n)
: (∃ S ⊆ X, S ≠ ∅ ∧ ∃ m : ℤ,
    |m + ∑ s in S, s| ≤ 1 / (n + 1))
```

Putnam 2006 B2

Putnam 2006 B2. Prove that, for every set $X = \{x_1, x_2, \ldots, x_n\}$ of n real numbers, there exists a nonempty subset S of X and an integer m such that

$$\left| m + \sum_{s \in S} s \right| \leq \frac{1}{n+1}.$$

```
(a) theorem putnam_2006_b2
(n : ℕ)
(npos : n > 0)
(X : Finset ℝ)
(hXcard : X.card = n)
: (∃ S ⊆ X, S ≠ ∅ ∧ ∃ m : ℤ,
    |m + ∑ s in S, s| ≤ 1 / (n + 1))
```

```
(b) theorem putnam_2006_b2:
fixes n :: nat
and X :: "real set"
assumes npos: "n > 0"
and hXcard: "finite X ∧ card X = n"
shows "∃ S ⊆ X. (S ≠ {}) ∧ (∃ m ::
     int.
     ¦m + (∑ s ∈ S. s)¦ ≤ 1 / (n + 1))"
```

Putnam 2006 B2

Putnam 2006 B2. Prove that, for every set $X = \{x_1, x_2, \ldots, x_n\}$ of n real numbers, there exists a nonempty subset S of X and an integer m such that

$$\left| m + \sum_{s \in S} s \right| \leq \frac{1}{n+1}.$$

```
(a) theorem putnam_2006_b2
(n : ℕ)
(npos : n > 0)
(X : Finset ℝ)
(hXcard : X.card = n)
: (∃ S ⊆ X, S ≠ ∅ ∧ ∃ m : ℤ,
    |m + ∑ s in S, s| ≤ 1 / (n + 1))
```

```
(b) theorem putnam_2006_b2:
fixes n :: nat
and X :: "real set"
assumes npos: "n > 0"
and hXcard: "finite X ∧ card X = n"
shows "∃ S ⊆ X. (S ≠ {}) ∧ (∃ m ::
     int.
     ¦m + (∑ s ∈ S. s)¦ ≤ 1 / (n + 1))"
```

```
(c) Theorem putnam_2006_b2
(n : nat)
(npos : gt n 0)
(X : list R)
(hXcard : length X = n)
: exists (presS: R -> Prop) (m: Z) (S: list R),
(neq (length S) 0) /\ (forall (x: R),
    In x S <-> (In x X /\ presS x))
/\ (Rabs (IZR m + (fold_left Rplus S 0))
    <= 1 / INR (n + 1)).
```

Evaluations

- **PutnamBench is hard**, no test methods solve >1% of problems.

PutnamBench: Lean	
Method	**Success Rate**
GPT-4	1/640
COPRA	1/640
ReProver (+r)	0/640
ReProver (−r)	0/640

PutnamBench: Isabelle	
Method	**Success Rate**
GPT-4	1/640
DSP	4/640
Sledgehammer	3/640

PutnamBench: Coq	
Method	**Success Rate**
GPT-4	1/417
COPRA	1/417
Tactician	0/417
CoqHammer	0/417

Conclusion

- We believe progress on PutnamBench will require significant breakthroughs in:
  1. Lemma Synthesis & Proof Planning
  2. Retrieval from & using existing formal maths libraries

https://trishullab.github.io/PutnamBench/