

# 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

<b>Benchmark</b>	<b>#</b>	<b>Natural Language</b>	<b>Lean</b>	<b>Isabelle</b>	<b>Coq</b>	<b>Factored Solution</b>
MINIF2F	488	✓	✓ <sup>†</sup>	✓		
PROOFNET	371	✓	✓ <sup>†</sup>			N/A
FIMO	149	✓	✓ <sup>†</sup>			
PUTNAMBENCH	640	✓	✓	✓	✓	✓

# Putnam 2006 B2

**Putnam 2006 B2.** Prove that, for every set  $X = \{x_1, x_2, \dots, x_n\}$  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  $X = \{x_1, x_2, \dots, x_n\}$  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)

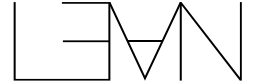
(npos : n > 0)

(X : Finset ℝ)

(hXcard : X.card = n)

: (∃ S ⊆ X, S ≠ ∅ ∧ ∃ m : ℤ,

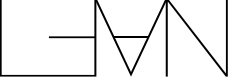
|m + ∑ s in S, s| ≤ 1 / (n + 1))



# Putnam 2006 B2

**Putnam 2006 B2.** Prove that, for every set  $X = \{x_1, x_2, \dots, x_n\}$  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   
( $n : \mathbb{N}$ )  
(npos :  $n > 0$ )  
( $X : \text{Finset } \mathbb{R}$ )  
(hXcard :  $X.\text{card} = n$ )  
: ( $\exists S \subseteq X, S \neq \emptyset \wedge \exists m : \mathbb{Z},$   
| $m + \sum s$  in  $S, s| \leq 1 / (n + 1)$ )

(b) **theorem** putnam\_2006\_b2:  
fixes  $n :: \text{nat}$   
and  $X :: \text{"real set"}$   
assumes npos: " $n > 0$ "  
and hXcard: " $\text{finite } X \wedge \text{card } X = n$ "  
shows " $\exists S \subseteq X. (S \neq \{\}) \wedge (\exists m ::$   
int.  
| $m + (\sum s \in S. s)| \leq 1 / (n + 1)$ )"





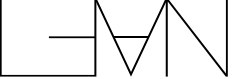
# Putnam 2006 B2


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

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

(b) **theorem** putnam\_2006\_b2:  
 fixes  $n :: \text{nat}$   
 and  $X :: \text{"real set"}$   
 assumes  $\text{npos: "n > 0"}$   
 and  $\text{hXcard: "finite X \wedge card X = n"}$   
 shows " $\exists S \subseteq X. (S \neq \{\}) \wedge (\exists m :: \text{int. } |m + (\sum s \in S. s)| \leq 1 / (n + 1))$ "



(a) **theorem** putnam\_2006\_b2   
 $(n : \mathbb{N})$   
 $(\text{npos} : n > 0)$   
 $(X : \text{Finset } \mathbb{R})$   
 $(\text{hXcard} : X.\text{card} = n)$   
 $: (\exists S \subseteq X, S \neq \emptyset \wedge \exists m : \mathbb{Z},$   
 $|m + \sum s \text{ in } S, s| \leq 1 / (n + 1))$

(c) **Theorem** putnam\_2006\_b2   
 $(n : \text{nat})$   
 $(\text{npos} : \text{gt } n \ 0)$   
 $(X : \text{list } R)$   
 $(\text{hXcard} : \text{length } X = n)$   
 $: \text{exists } (\text{presS} : R \rightarrow \text{Prop}) (m : \mathbb{Z}) (S : \text{list } R),$   
 $(\text{neq } (\text{length } S) \ 0) \wedge (\text{forall } (x : R),$   
 $\text{In } x \ S \leftrightarrow (\text{In } x \ X \wedge \text{presS } x))$   
 $\wedge (R.\text{abs } (\text{IZR } m + (\text{fold\_left } R.\text{plus } S \ 0))$   
 $\leq 1 / \text{INR } (n + 1)).$

# 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/>