

Formally proving the correctness of complex protocols and systems with TLA+ Proof System (TLAPS)

<https://github.com/specula-org/TLAPS-Bench>

Overview and Value Proposition

A service for verifying complex protocols and systems in TLA+

- It is used to verify real-world systems (together with Specula)

A benchmark for challenging frontier AI in writing formal proofs

- Frontier AI (e.g., Astra) can only solve less than 50% of the proof tasks.

TLAPS-Bench 2.0

- **9 real-world TLA+ specs**
- **56 invariants**

Problem	Type	#Spec	#Inv
FLASH Cache coherence	Protocol	1	15
ZooKeeper protocol	Protocol	1	9
Cahill's snapshot isolation	Protocol	1	1
Ivy TLB	System	1	2
Open Addressing	System	1	5
EtcD Raft	System	1	8
HashiCorp Raft	System	1	6
ZooKeeper code	System	1	9
MongoDB dist. transaction	System	1	1

* TLAPS-Bench has **123 specs and 301 invs** in total

* Problem sets are actively evolved to challenge AI

TLAPS-Bench Runtime

TLAPS-Bench provides a runtime for agents to do proof tasks

- Equipping agents with tools, e.g., SANY, TLC, and Apalache
- Anti-cheating features
- Sandboxing and isolation

We have used the TLAPS-Bench runtime to prove many specs

- TCP state machine, CFT and BFT protocols, and many others

Results

Model (Agent)	Success Rate	Total Hours	Tokens In / Out per Inv. (M)	Cost	Cost per Inv.
 GPT-6 Astra, max (Codex)	41.1% (23/56)	-	10.0 / 0.027	\$653.36	\$11.67
 Opus 5, max (Claude Code)	37.5% (21/56)	58	16.4 / 0.092	\$1,319	\$24
 Muse Spark 1.3, xhigh (OpenCode, contributor)	3.6% (2/56)	24.5	12.9 / 0.057	\$4.00	\$0.07

- The problems are **hard for frontier AI** (we have many harder problems)
- The proof tasks are **long-horizon**, taking hours (or days) to finish.