

Summary of LeanDojo: Theorem-Proving Infrastructure for Lean

1 Overview of the Paper

LeanDojo is a unified infrastructure designed to support formal theorem proving in the Lean 4 proof assistant. The paper identifies the need for more effective and scalable machine learning methods that can interact with formal proof systems like Lean. To this end, the authors develop LeanDojo to bridge the gap between ML and proof engineering.

2 Key Components of LeanDojo

2.1 Unified Interface

LeanDojo provides a unified Python interface to access and manipulate Lean projects using Lean 4 APIs. This includes support for:

- Parsing and indexing Lean files.
- Extracting and modifying goals, environments, and contexts.
- Tracing proof states and tactic invocations.

2.2 Dataset Creation

The authors construct a large dataset of Lean proofs from existing projects like Mathlib, covering over 100,000 tactic steps. The dataset includes:

- Goal states.
- Local context and environment.
- Tactics and their effects.

2.3 LeanGym Environment

LeanDojo includes LeanGym, an OpenAI Gym-like environment for reinforcement learning in Lean. It simulates:

- Actions: tactic invocations.
- States: Lean proof goals and contexts.
- Rewards: proof progress and final success.

2.4 Tactic Prediction Models

The paper benchmarks baseline models (transformers, seq2seq) for predicting next-step tactics, using the LeanDojo dataset. These models demonstrate the feasibility of data-driven Lean automation.

3 Application in Our Project

We can leverage LeanDojo in our IPO compliance project in the following ways:

3.1 Automated Rule Tracing and Verification

Use LeanDojo’s APIs to trace goal states and tactic outcomes across modular rule definitions like ‘rule_5.1.b’, enabling explainability.

3.2 LLM-HtT Prompting + LeanDojo

Use LLMs to convert natural language rule text to Lean predicates, then use LeanDojo to verify syntactic correctness and proof compatibility.

3.3 Reinforcement Learning for Proof Search

Use LeanGym to experiment with RL agents that learn to compose rule predicates or explore logical alternatives for eligibility conditions.

3.4 Lean Proof Dataset from Domain

Extract Lean proof snippets from structured regulatory logic and red herring filings to build a domain-specific Lean dataset.

4 Future Steps

Step 1 Integrate LeanDojo API: Build a wrapper over our Lean rules to log proof states.

Step 2 Modular Rule Library: Refactor all rules as individually provable predicates with named identifiers.

Step 3 LLM-to-Lean Verification: Use LeanDojo to validate LLM-translated rules and hypotheses.

Step 4 Proof Audit Logging: Store proof paths, failure reasons, and contexts to support audit trails.

Step 5 Explore LeanGym: Prototype an agent for structured proof search across eligibility conditions.

5 Conclusion

LeanDojo provides a powerful infrastructure for bridging ML and formal verification. By adopting its tooling, we can make our IPO compliance checker more robust, explainable, and scalable, ultimately pushing towards a trusted AI+Law product.