Agent-Based Automated Proof Engineering in Real Formal Libraries

Wenda Li

Recent advances in AI have led to strong performance in isolated theorem-proving tasks. However, maintaining large formal libraries involves sustained proof engineering efforts, including library extension, proof refactoring, and adaptation to evolving dependencies. These activities differ substantially from single-problem theorem proving and remain largely unaddressed by existing benchmarks and systems.

We propose Automated Proof Engineering (APE) as a research setting that targets realistic proof maintenance tasks arising in real formal developments. As a concrete instantiation, we introduce APE-Bench, a benchmark constructed from actual library commits that captures a wide range of proof engineering scenarios. To address the heterogeneity of these tasks, we present APE-Agent, a unified agent framework that supports autonomous behavior across diverse formal mathematics activities within a single architecture.

Beyond formal mathematics, we argue for a paradigm shift from prevailing Generative AI approaches, which center on producing plausible standalone outputs. We instead advocate a paradigm in which models actively leverage verification mechanisms to overcome uncertainty in both their environment and their own knowledge, and refer to this paradigm as Certified AI.