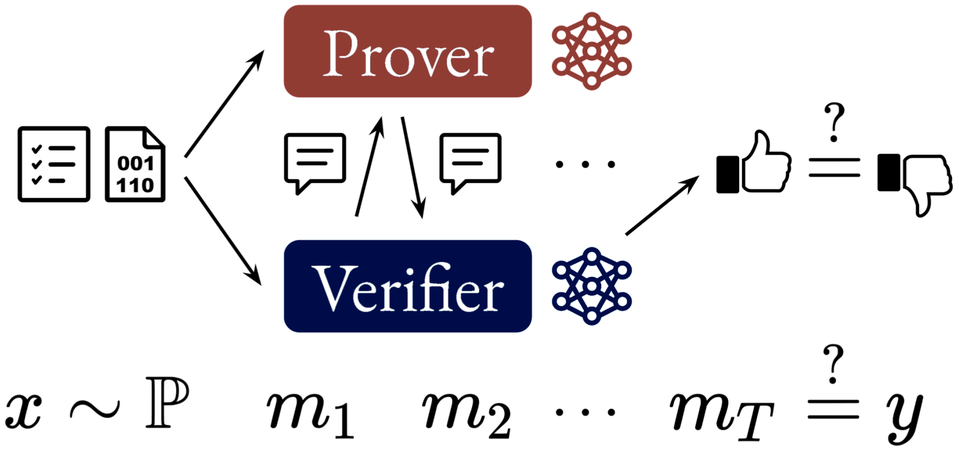
Neural Interactive Proofs

Lewis Hammond and Sam Adam-Day

Department of Computer Science
University of Oxford

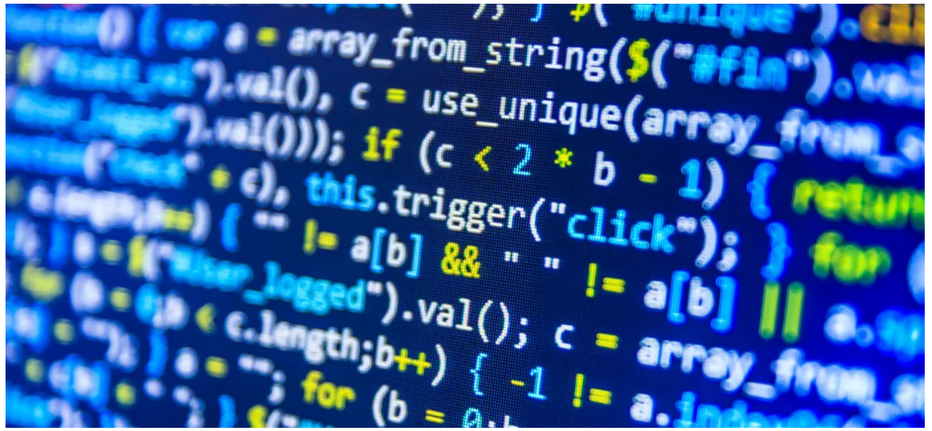
Overview

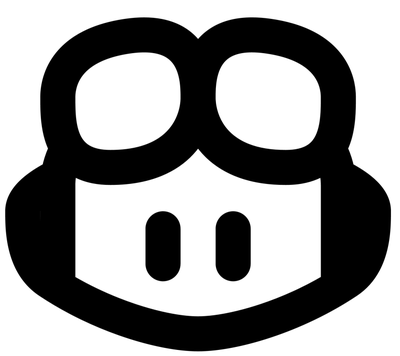
- We study how a weak but trustworthy verifier can learn to interact with a strong but untrusted prover in order to solve tasks beyond the verifier's capabilities
- In particular we represent provers and verifiers using neural networks and formalise their interactions using prover-verifier games (Anil et al., 2021)
- In particular, we provide:
  - Theoretical motivation
  - New interaction protocols
  - Empirical comparison
  - A comprehensive codebase

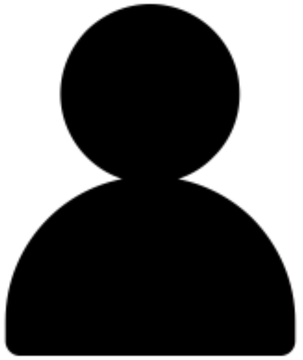


Motivation

- We have large, powerful machine learning models, but it can be difficult to trust them
- Traditional approaches to verification don't scale to state-of-the-art models
- Can we use weak, trusted agents to verify stronger, untrusted agents?
  - E.g. training a verifier agent to analyse the code produced by a prover agent







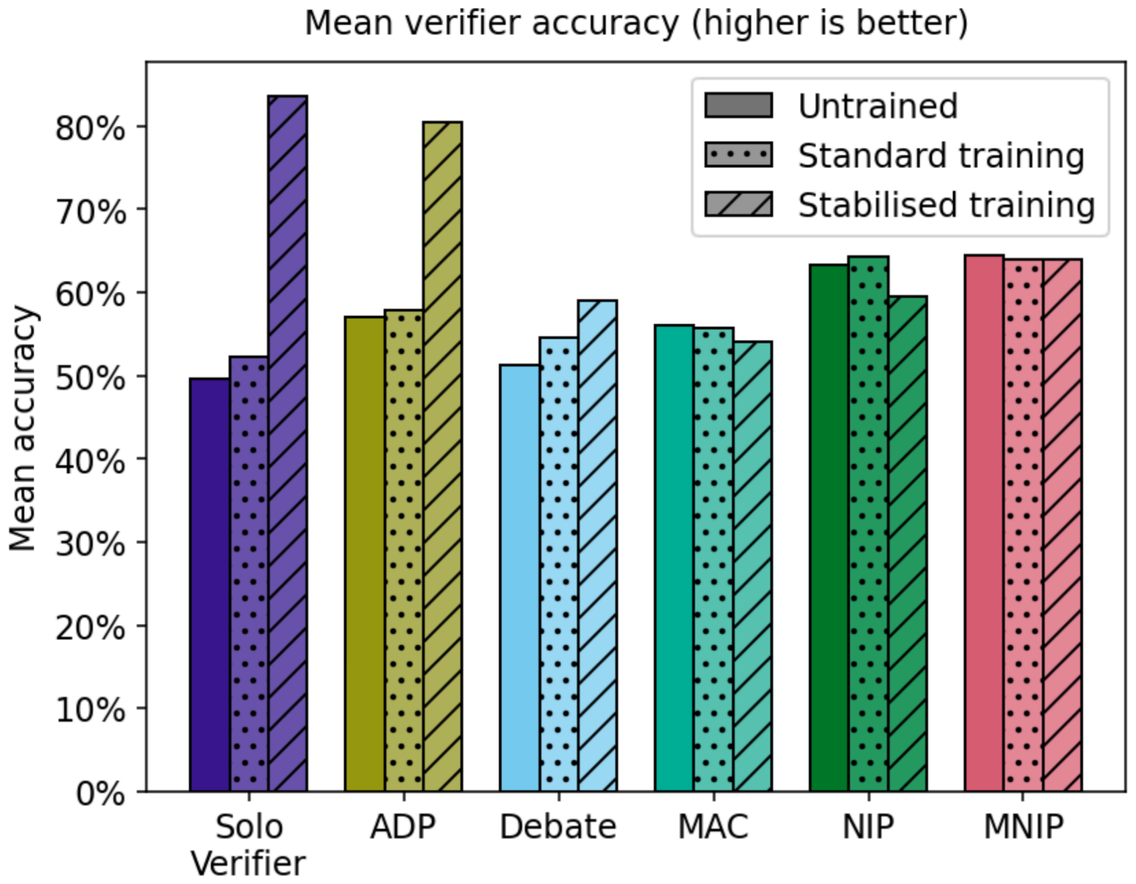
Theoretical Results

- We provide equivalences between the equilibria of the different prover-verifier games and valid proof systems
- Reaching these equilibria is challenging, however, as they represent Stackelberg equilibria over worst-case losses

Protocol	Provers	Verifiers	Rounds	Complexity	ZK	Reference
`adp`	1	1	2	NP	✗	(Anil et al., 2021)
`debate`	2	1	T	PSPACE	✗	(Irving et al., 2018)
`mac`	2	1	2	MA	✗	(Wäldchen et al., 2024)
`nip`	1	1	T	PSPACE	✗	Ours
`mnip`	2	1	T	NEXP	✗	Ours
`zk-nip`	1	3	T	PSPACE	✓	Ours
`zk-mnip`	2	3	T	NEXP	✓	Ours

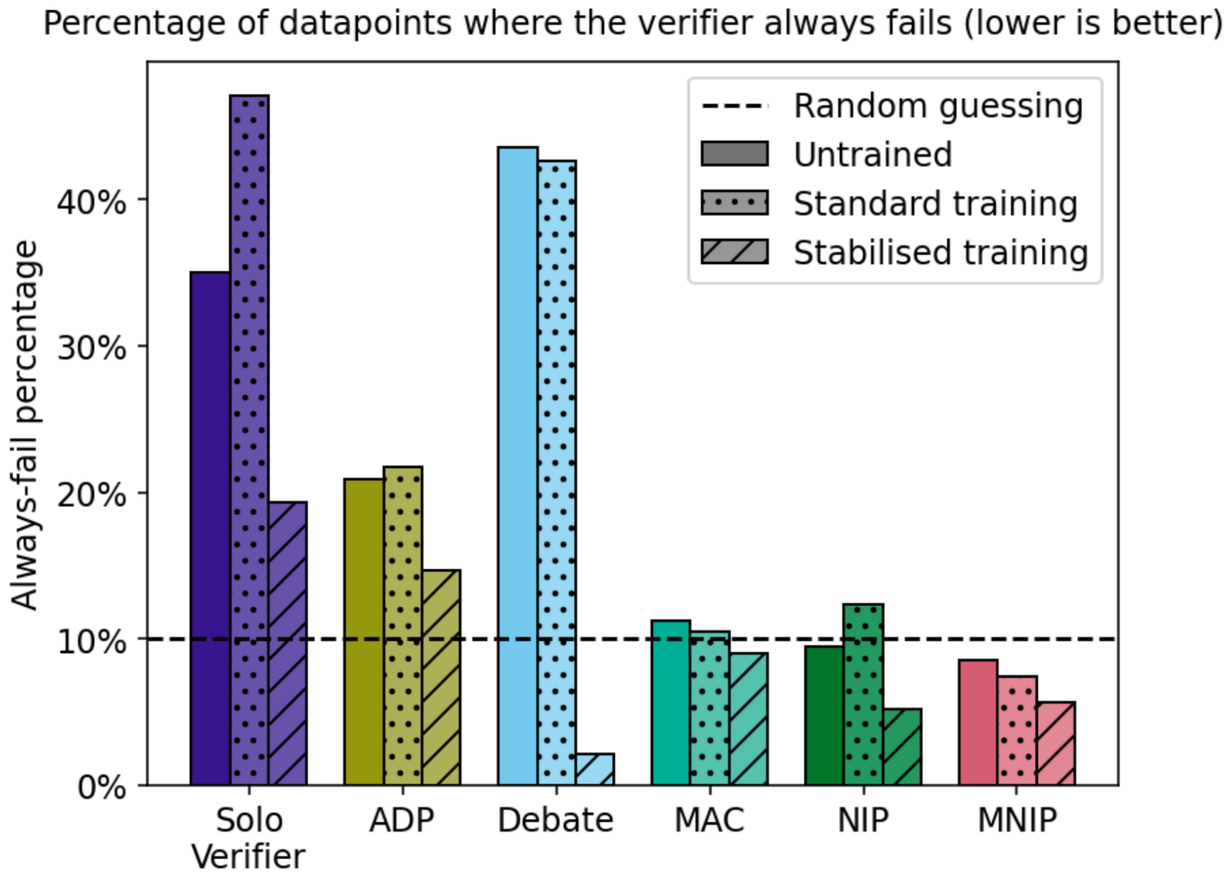
Experimental Results

- We compare our new protocols against existing proposals in two domains:
  - Graph isomorphism
  - Code validation
- We find that:
  - Our new protocols perform favourably (and more robustly)
  - But there is still much progress to be made on training algorithms



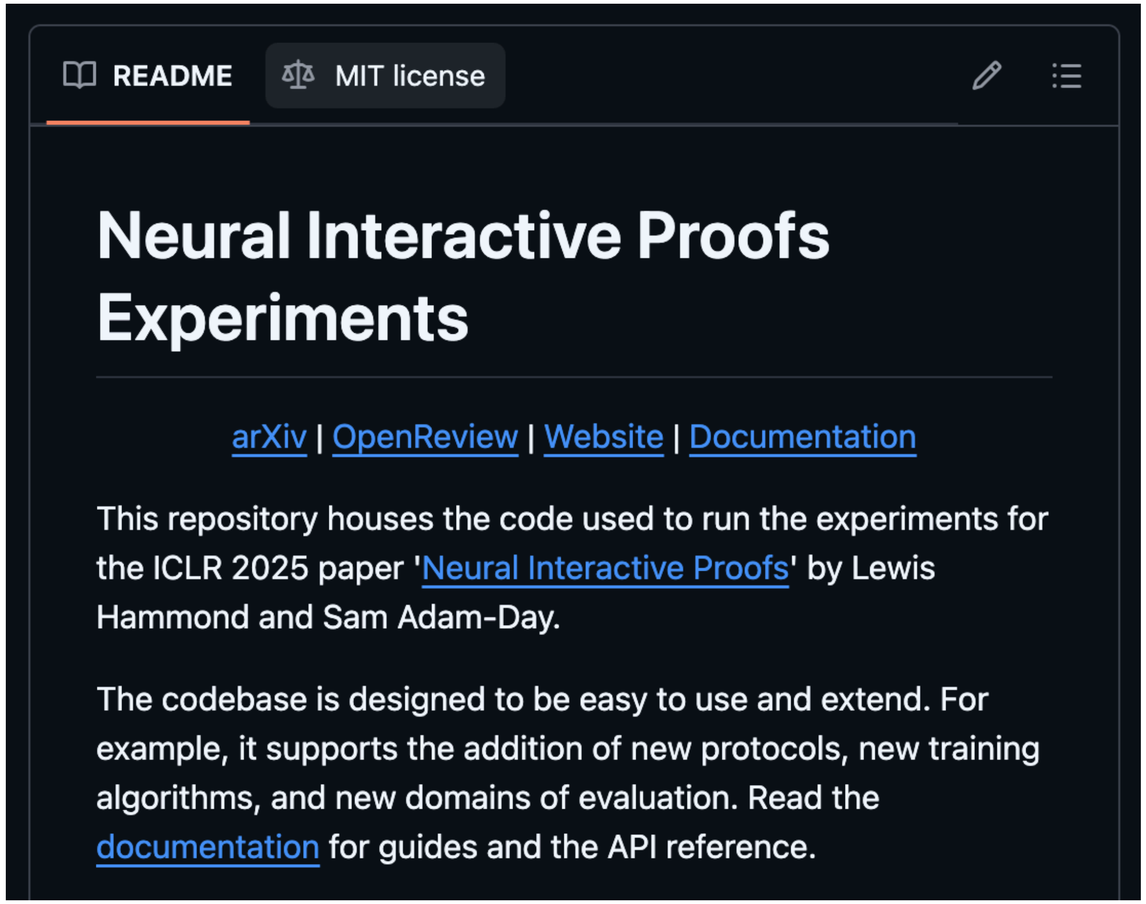
Experimental Results

- We compare our new protocols against existing proposals in two domains:
  - Graph isomorphism
  - Code validation
- We find that:
  - Our new protocols perform favourably (and more robustly)
  - But there is still much progress to be made on training algorithms



Next Steps

- In future, we plan to test:
  - New training algorithms
  - New domains
  - New interaction protocols
  - The effects of scale
- Our theoretical and empirical contributions aim to provide a foundation for future work on these topics



Thanks for listening!

For the paper, code, and more, go to neural-interactive-proofs.com