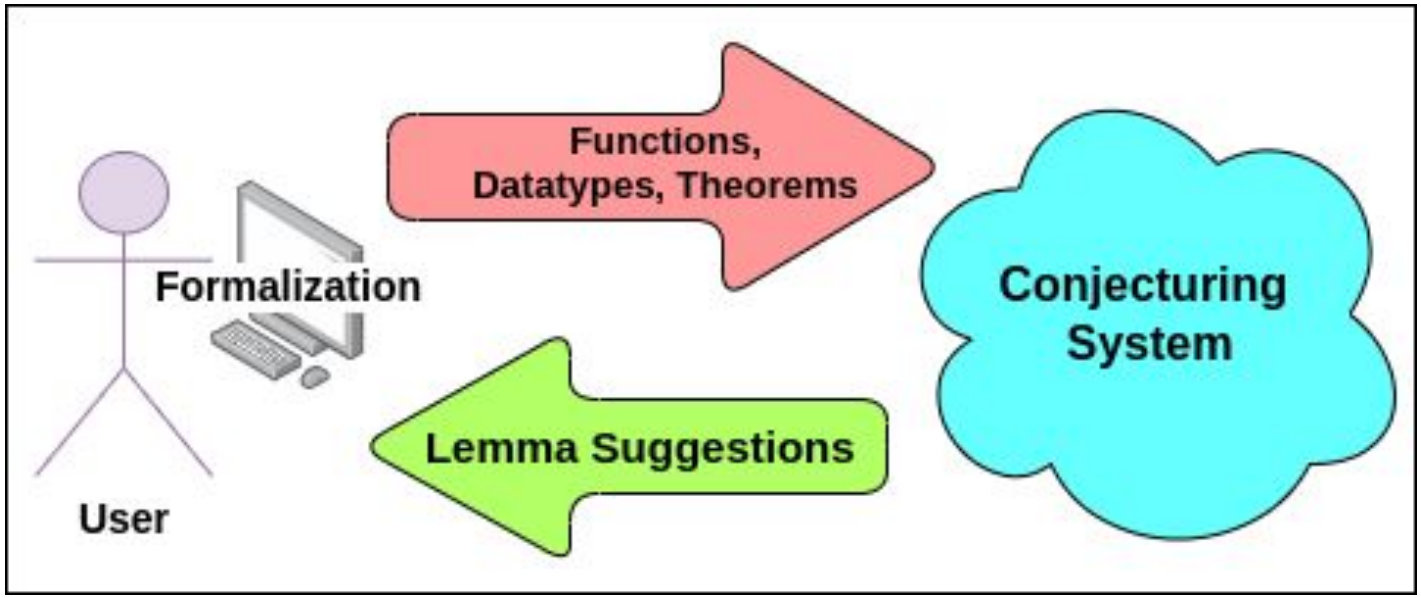
Lemmanaid:
Neuro-Symbolic Lemma Conjecturing

Sólrún Halla Einarsdóttir

In collaboration with:
Yousef Alhessi, Emily First, George Granberry,
Moa Johansson, Sorin Lerner, Nick Smallbone

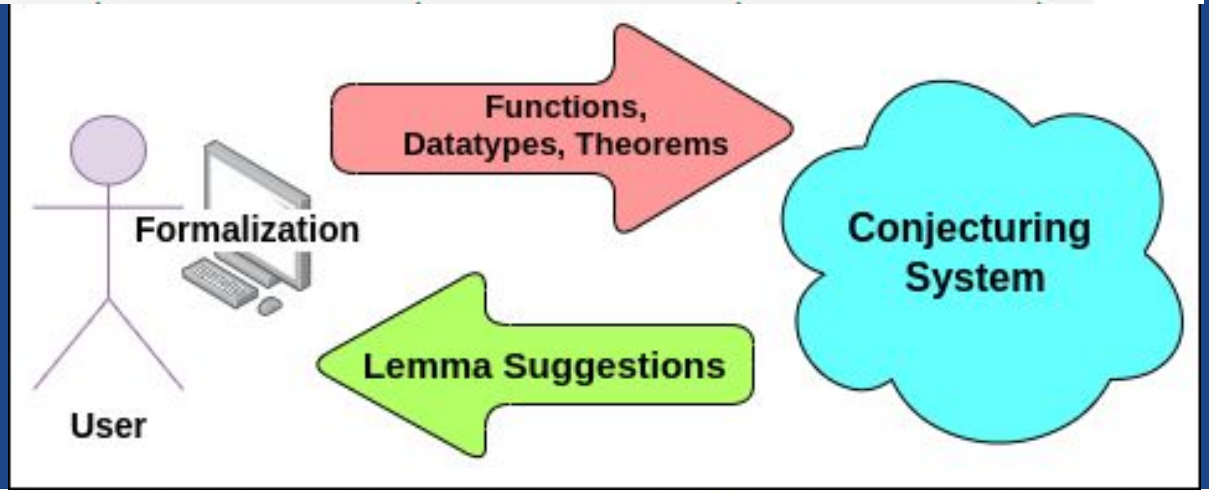
AITP 2025

Conjecturing



```
codatatype octo =
  Octo (Ree: real) (Im1: real) (Im2: real) (Im3: real) (Im4: real)
       (Im5: real) (Im6: real) (Im7: real)
```

```
definition times_octo :: "[octo, octo] ⇒ octo"
  where
    "(a * b) =  (let   ...
```




```
lemma octo_product_noncommutative: "¬(∀x y::octo. (x * y = y * x))"
```

```
lemma octo_distrib_left :
  "a * (b + c) = a * b + a * c" for a b c ::octo
```

Example - Octonions - Lemma Templates

```
lemma octo_product_noncommutative: "¬(∀x y::octo. (x * y = y * x))"
```

```
¬ (∀y_0 y_1. ?H1 y_0 y_1 = ?H1 y_1 y_0)
```

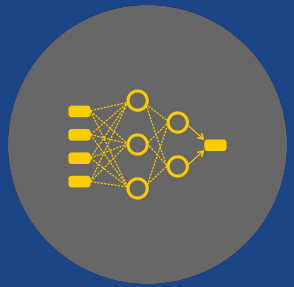
```
lemma octo_distrib_left :
  "a * (b + c) = a * b + a * c" for a b c ::octo
```

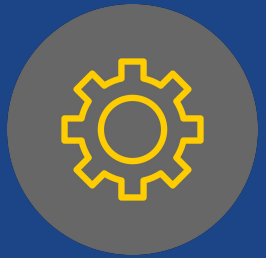
```
?H1 x_1 (?H2 x_2 x_3) = ?H2 (?H1 x_1 x_2) (?H1 x_1 x_3)'
```

Insight

Use LLM to generate templates

Use template-directed synthesizer to generate lemmas

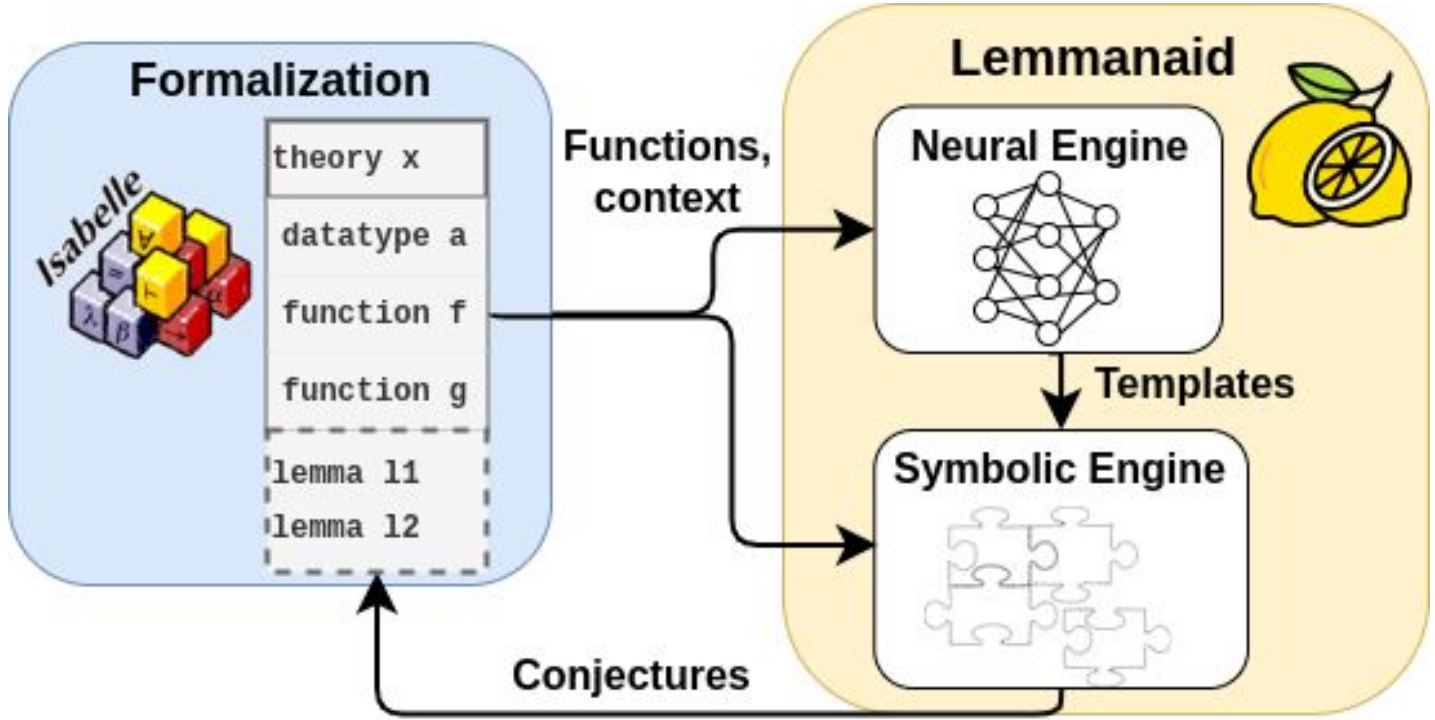




Language Model (neural)

Synthesizer (symbolic)

Lemmanaid for template-based conjecturing



Evaluation - Datasets

HOL Library

(Homogenous Dataset)

62,815 Theorems

- Train on 57,575
- Eval on 4,740

Archive of Formal Proofs (AFP)

(Diverse Dataset)

206,304 Theorems

- Train on 189,442 (prior to 2024)
- Eval on 16,362 (added in 2024)

Octonions

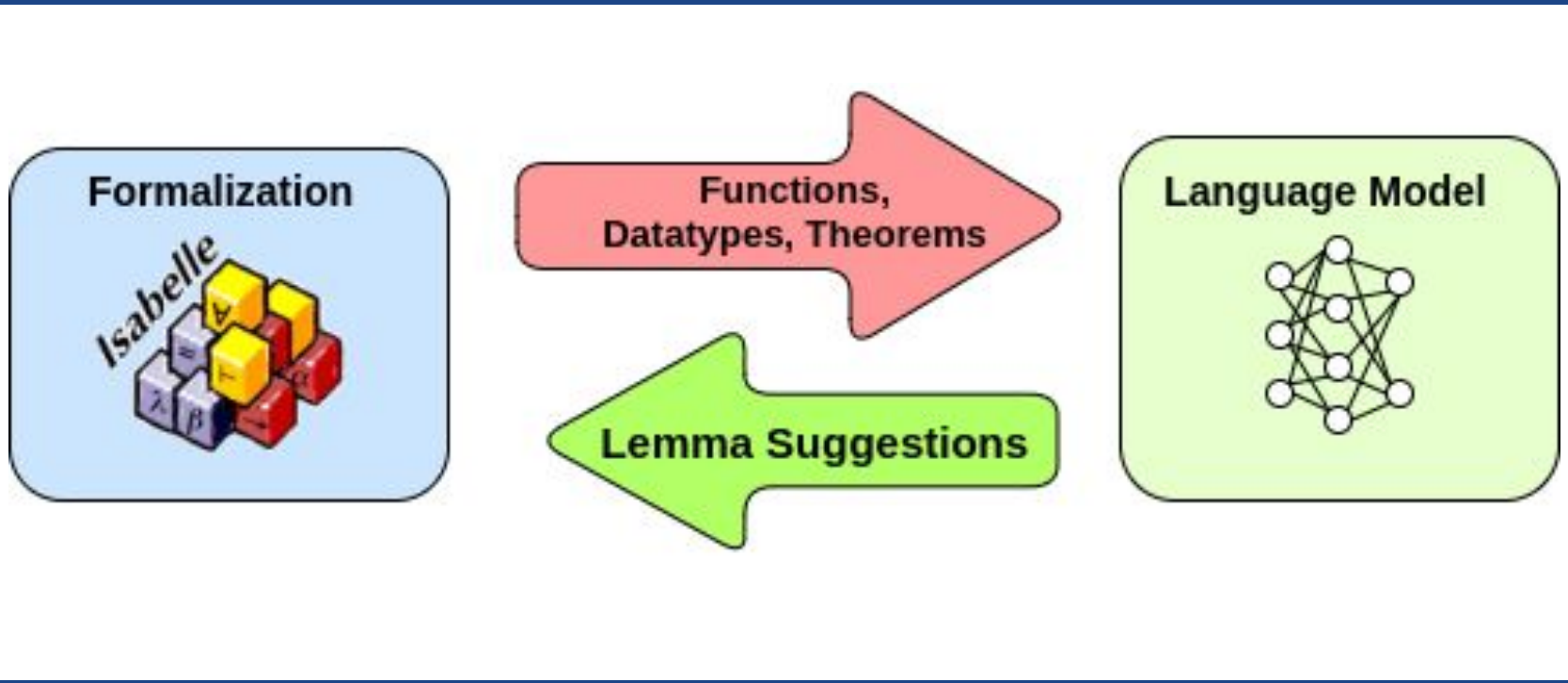
(Case Study)

350 Theorems

For eval only

Enables comparison to symbolic tool

Neural Conjecturing



Test-Time Compute

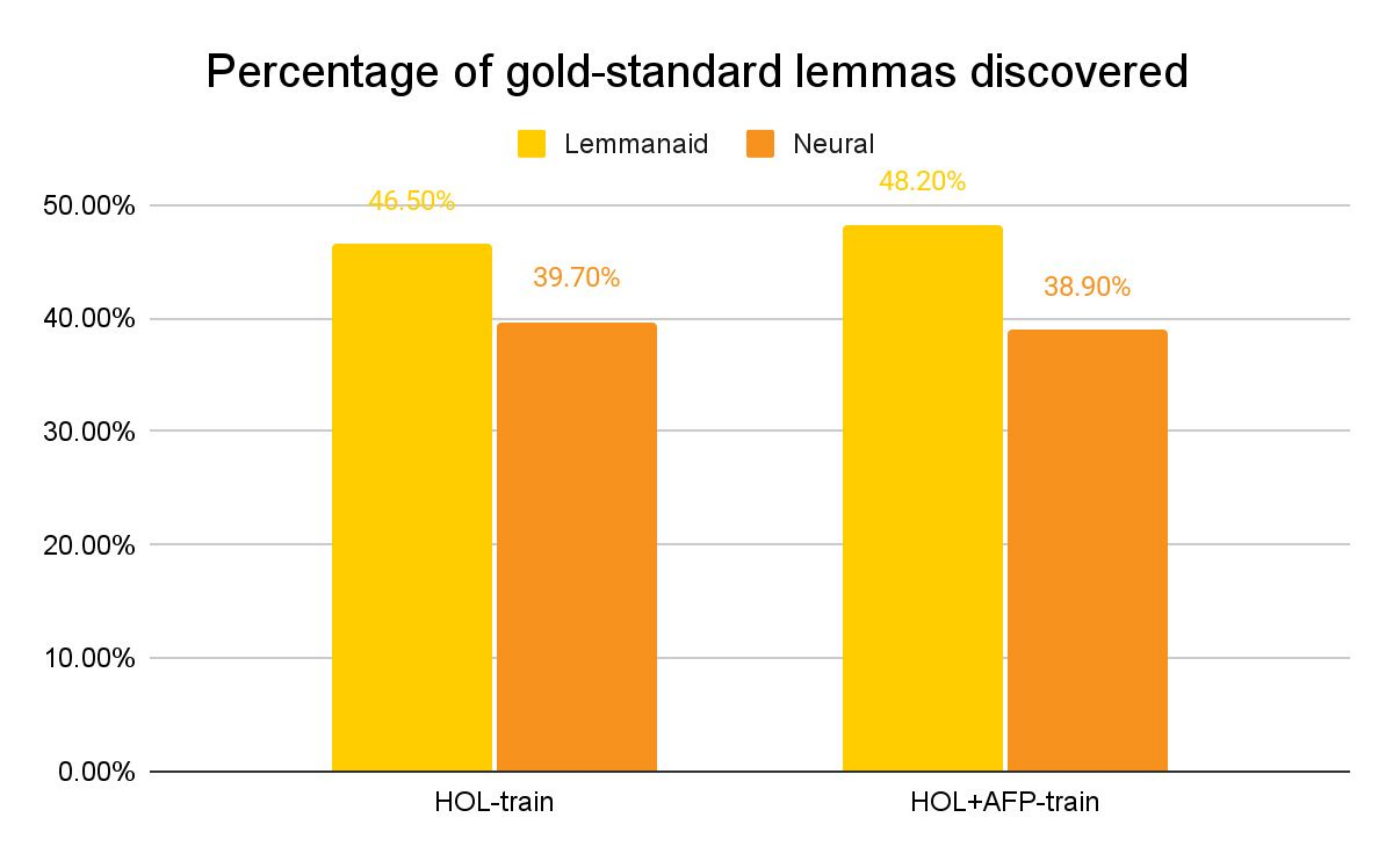
We sample 5 predictions from each model:

- 1 greedy-decoding
- 4 beam-search

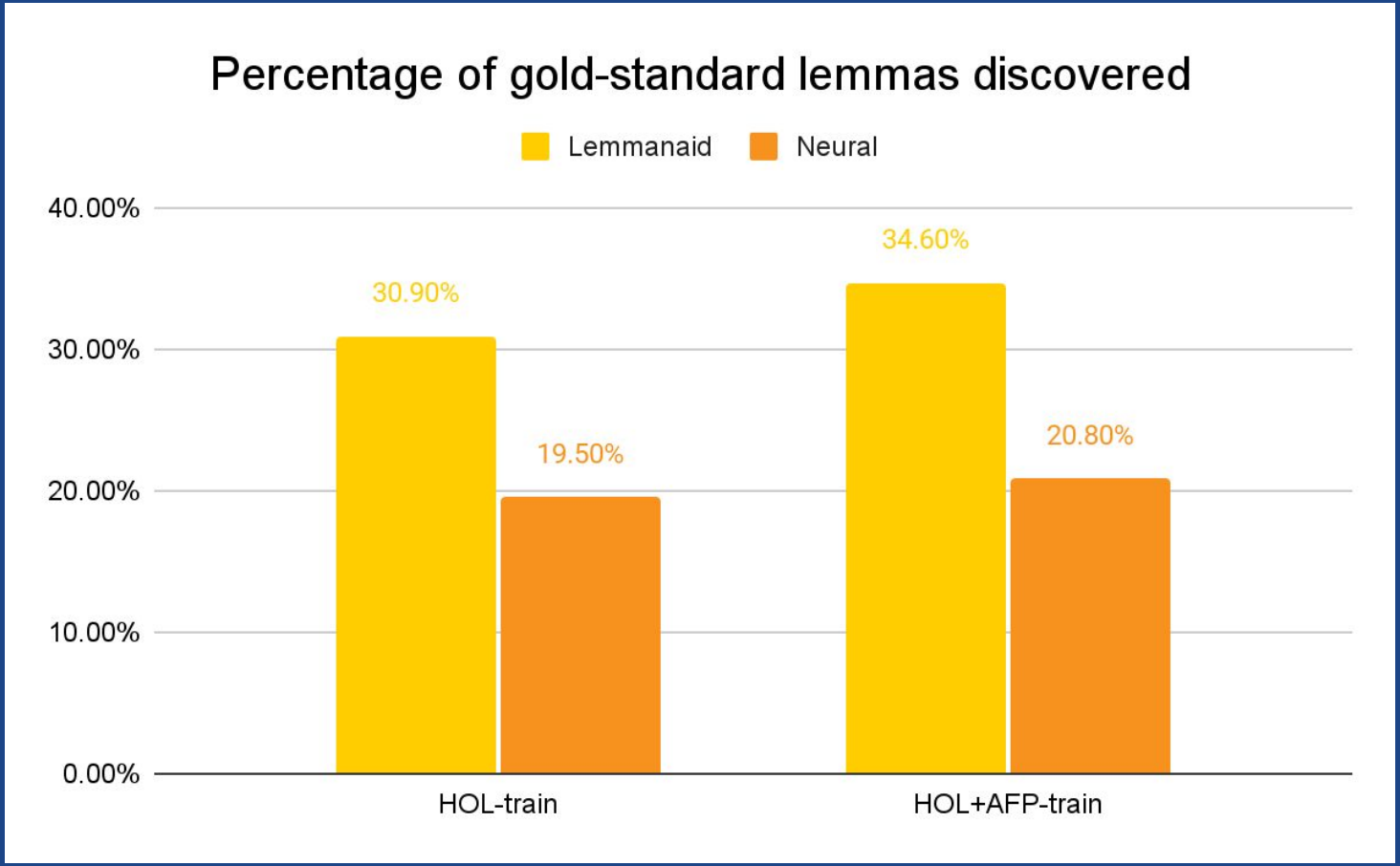
A prediction is correct **if it matches target lemma**

- Lemma defined and proved by human expert is probably interesting/useful
- Likely under-reporting results for both models

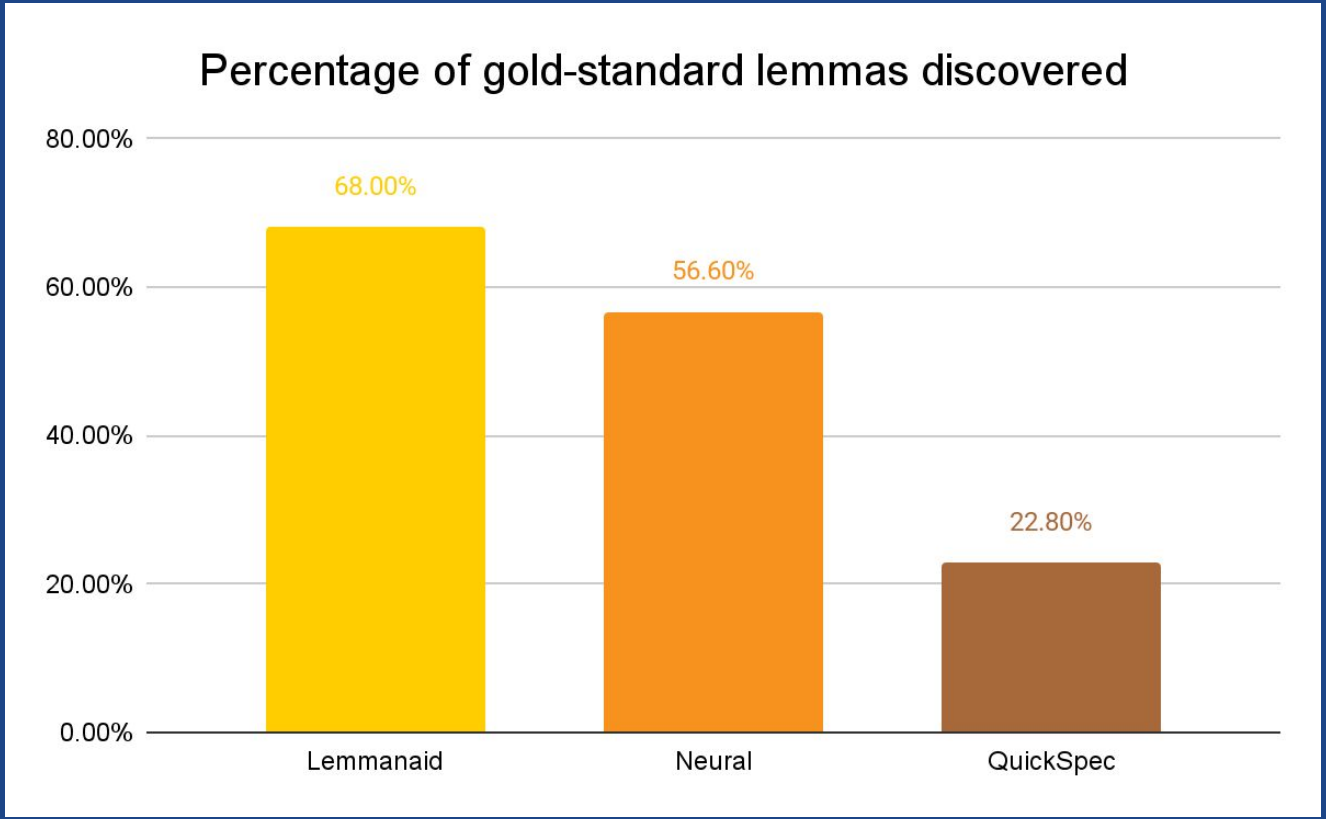
Results - HOL Library



Results - AFP test



Results - Octonions



Conclusions + Ongoing work

- Neuro-symbolic conjecturing outperforms both purely neural and purely symbolic
- Ongoing work:
  - Extend evaluation of results - could a generated lemma be valid and useful while not matching a given target?
  - Try bigger models! (Current results were found using DeepSeek-coder-1.3b)

Check out our preprint:
https://arxiv.org/abs/2504.04942