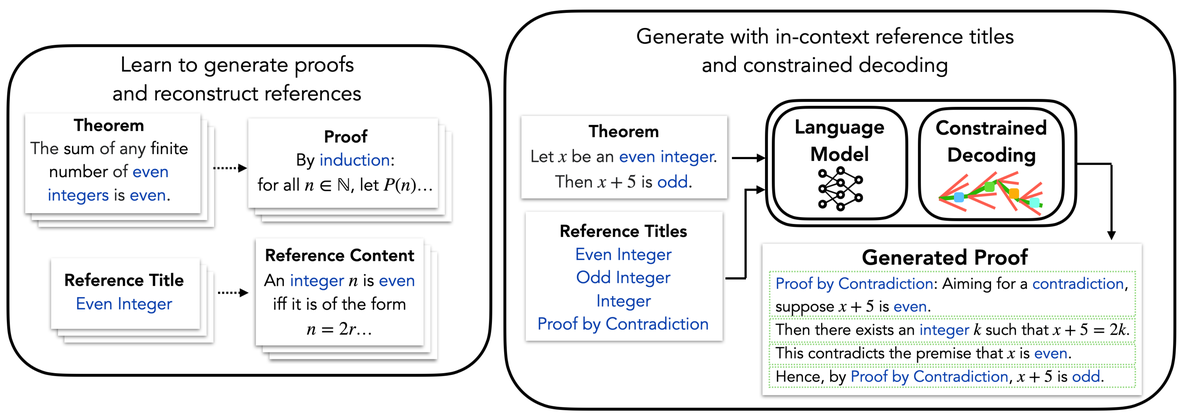
NaturalProver: Grounded Mathematical Proof Generation with Language Models

Sean Welleck*[1,2] Jiacheng Liu*[1] Ximing Lu[1,2] Hannaneh Hajishirzi[1,2] Yejin Choi[1,2]

[1]Paul G. Allen School of Computer Science, University of Washington [2]Allen Institute for Artificial Intelligence

NaturalProver

Can large language models **help people prove mathematical theorems**?



We present **NaturalProver**, a language model that generates mathematical proofs by conditioning on background references (e.g. theorems and definitions that are either retrieved or human-provided), and optionally enforces their presence with constrained decoding.

Grounding = references + constrained decoding

NaturalProver is an instance of GPT-3 fine-tuned on NaturalProofs [Welleck et al., Neurips 2021]. NaturalProver adds two components on top of GPT-3:

- **In-context references**: retrieved or provided theorems/definitions relevant to a correct proof.
- **Constrained decoding**: samples multiple next-steps, retains steps in a beam based on constraints.

Natural vs. formal theorem proving

```
theorem aime_1984_p1
  (u : ℕ → ℚ)
  (h₀ : ∀ n, u (n + 1) = u n + 1)
  (h₁ : ∑ k in finset.range 98, u k.succ = 137) :
  ∑ k in finset.range 49, u (2 * k.succ) = 93 :=
begin
  rw finset.sum_eq_multiset_sum,
  dsimp [finset.range] at h₁,
  simp [h₀],
  ring,
  norm_num at h₁,
  norm_num,
  apply eq_of_sub_eq_zero,
  { simp only [*, abs_of_pos, add_zero] at *, linarith },
end
```

- Rigid
- Not much data
- Easy to verify

Theorem:
Let x be an even integer.Then $x + 5$ is odd.

Proof by Contradiction: Aiming for a contradiction, suppose $x + 5$ is even.
Then there exists an integer k such that $x + 5 = 2k$.
This contradicts the premise that x is even.
Hence, by Proof by Contradiction, $x + 5$ is odd.

- Flexible
- Used in education, science, engineering
- Lots of language data
- Hard to verify!

Figure 1. Classical provers use **rigid formal languages.** Can LLMs prove in **flexible natural language**?

Capable of correct, useful proof generation

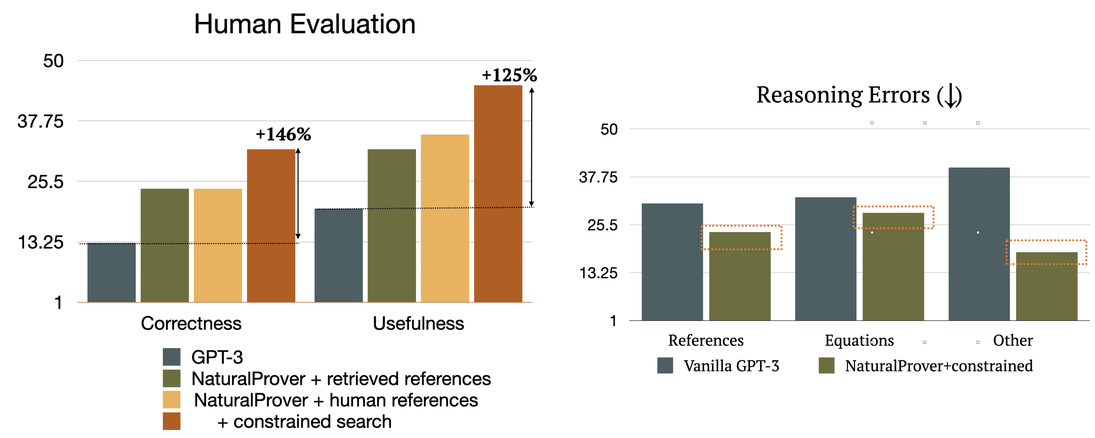


Figure 2. On theorems from the NaturalProofs benchmark, NaturalProver improves the quality of next-step suggestions and generated proofs over fine-tuned GPT-3, according to human evaluations from university-level mathematics students. NaturalProver is capable of proving some theorems that require short (2-6 step) proofs, and providing next-step suggestions that are rated as correct and useful over 40% of the time.

Human-machine collaboration

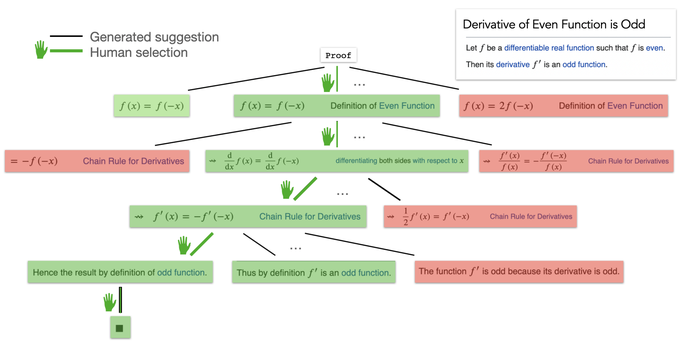


Figure 3. NaturalProver had > 40% correct and useful *next-step predictions*. These compound in full-proof generation. An exciting option is ***human-machine collaboration*** with *multiple* suggestions.

Towards verified natural proofs : come see us at the *MathAI workshop*!