Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs

Albert Q. Jiang, Sean Welleck, Jin Peng Zhou, Timothée Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, Yuhuai Wu

Paper: Dataset:

NEURAL INFORMATION PROCESSING SYSTEMS

The divide between informal and formal mathematics

Informal mathematics

- Reasoning with flexibility
- Abundant data
- Flexible reasoning
- Verification in limited circumstances
- Prone to error and false positives

Formal mathematics

- Reasoning with rigour
- Signal in the middle of a proof
- Can potentially verify all mathematical domains
- Limited data
- Fairly rigid reasoning with still not-so-perfect automation

The best of both worlds

Reason informally, and prove formally.

- Humans are extremely good at informal reasoning (though imperfect).
- Language models (Minerva) have also shown impressive informal mathematical reasoning capabilities.

Drafting informal solutions

- Humans are extremely good at informal reasoning (though imperfect).
- Language models (Minerva [1]) have also shown impressive informal mathematical reasoning capabilities.

Sketching with few-shot learning

A formal sketch is a sequence of formal conjectures expressing the high-level ideas of the proof. It is well-aligned with the informal proof.

- Codex input:
  - Informal statement 1
  - Informal proof 1
  - Formal statement 1
  - Formal sketch 1
  - Informal statement 2
  - Informal proof 2
  - Formal statement 2
  - Formal sketch 2
  - Informal statement 3
  - Informal proof 3
  - Formal statement 3
- Codex output:
  - Formal sketch 3

Informal Statement: Show that for any real number a, $10a \leq 28a^2 + 1$.

Informal Proof:
It suffices to show $0 \leq 28a^2 - 10a + 1$. First, consider completing the square for $28a^2 - 10a$ and observe that $\left(a - \frac{5}{28}\right)^2 = a^2 - \frac{10}{28}a + (5/28)^2$. Since $0 \leq \left(a - \frac{5}{28}\right)^2$, we get $0 \leq a^2 - \frac{10}{28}a + (5/28)^2$. Multiplying by 28 and simplifying gives $0 \leq 28a^2 - 10a + (25/28)$. Since $25/28 < 1$, the result follows.

Formal Statement:

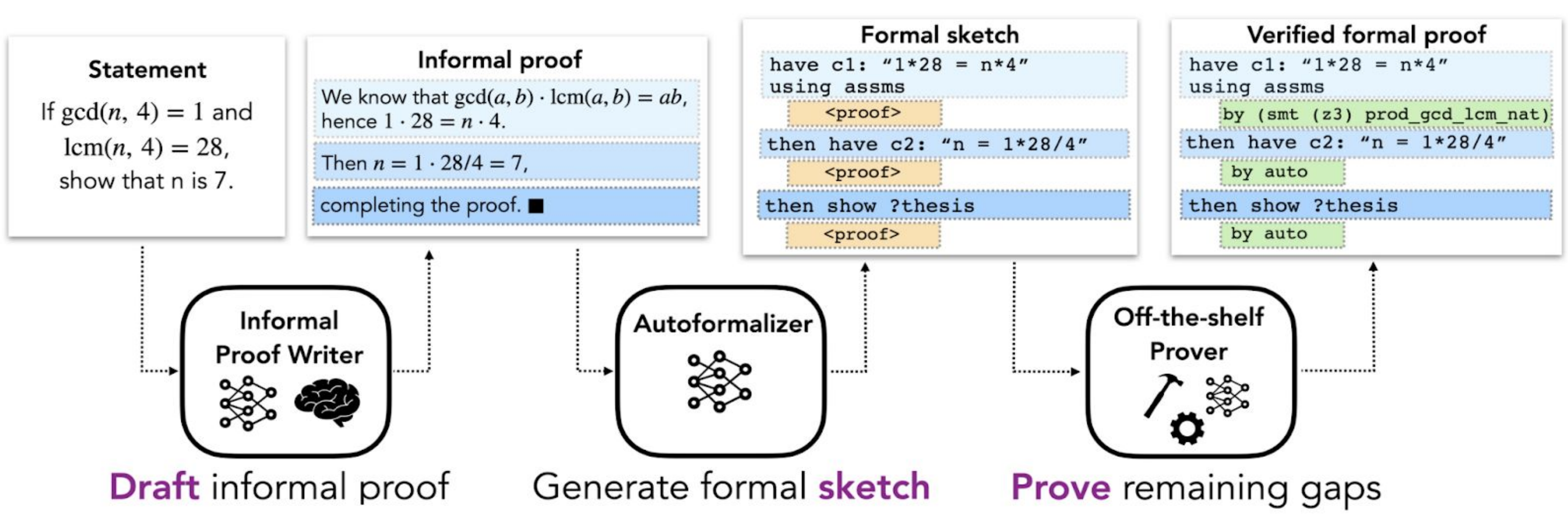
```
theorem algebra_binomnegdiscrineq_10alt28asqp1:
  fixes a :: real
  shows "10 * a ≤ 28 * a^2 + 1"
```

Formal Proof Sketch:

```
proof - (* it suffices to show 0 <= 28a^2 - 10a + 1 *)
  have c0: "0 ≤ 28a^2 - 10a + 1"
  proof - (* observe that (a - (5/28))^2 = a^2 - (10/28)a + (5/28)^2 *)
    have c1: "(a - (5/28))^2 = a^2 - 10/28a + (5/28)^2" <···>
    (* we get 0 <= a^2 - (10/28)a + (5/28)^2 *)
    have c2: "0 ≤ a^2 - 10/28a + (5/28)^2" using c1 <···>
    (* Multiplying by 28 and simplifying gives 0 <= 28a^2 - 10a + (25/28) *)
    have c3: "0 ≤ 28a^2 - 10a + 28((5/28)^2)" using c2 <···>
    have c4: "0 ≤ 28a^2 - 10a + 28((5/28)*(5/28))" using c3 <···>
    have c5: "0 ≤ 28a^2 - 10a + (25/28)" using c4 <···>
    (* Since 25/28 < 1, the result follows. *)
    show ?thesis using c5 <···>
  qed
  show ?thesis <···>
qed
```

Proving open conjectures in the sketches

- To verify the correctness of the formal sketches, we need to close the "gaps" in them.
- We use a symbolic automated theorem proving tool (Sledgehammer + heuristics), but in principle any off-the-shelf prover can be used.



The Draft, Sketch, and Prove (DSP) process illustrated

Experimental results

- We experiment on the miniF2F dataset [2], a collection of 488 high-school competition level mathematical problems.
- It is divided into a validation and a test set, but we do not differentiate them in this work.
- We generate 100 informal proofs from each language model and sketch once per proof.

Success rate	miniF2F-valid	miniF2F-test
Baselines		
Sledgehammer	9.9%	10.4%
Sledgehammer + heuristics	18.0%	20.9%
Thor (Jiang et al., 2022)	28.3%	29.9%
Thor + expert iteration (Wu et al., 2022)	37.3%	35.2%
Draft, Sketch, and Prove		
Human informal proof	42.6%	**39.3%**
Codex informal proof	40.6%	35.3%
8B Minerva informal proof	40.6%	35.3%
62B Minerva informal proof	**43.9%**	37.7%
540B Minerva informal proof	42.6%	38.9%

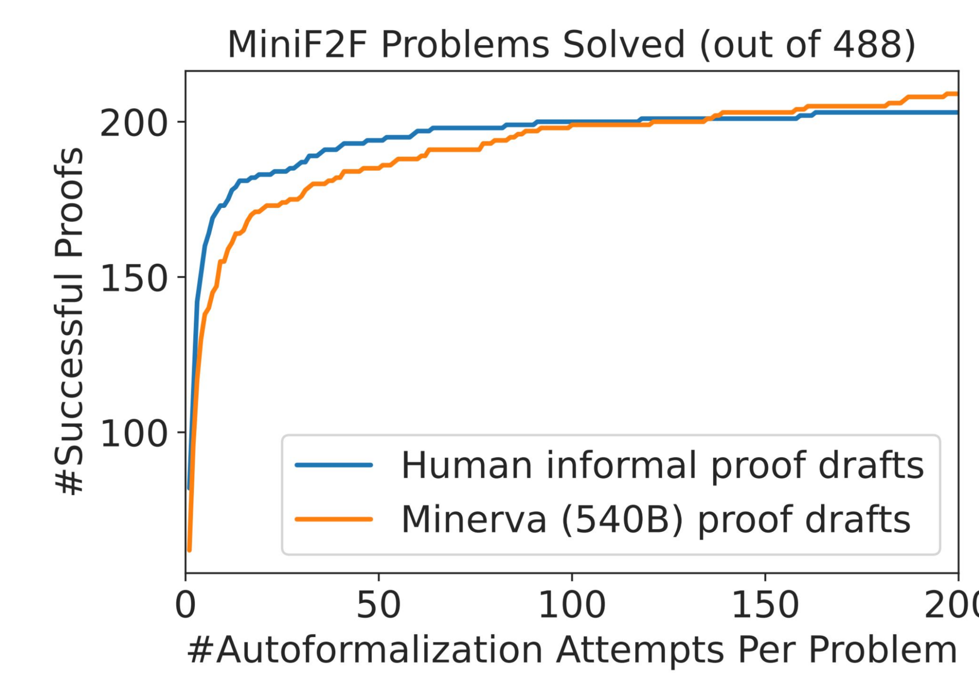
Previous SOTA

Best performance on test

Best performance on valid

The performance of Draft, Sketch, and Prove with various sources of informal proofs, and baseline methods with Isabelle.

- DSP almost doubles the automated prover's performance.



- Language model proof drafts close more problems than human ground truths??!!🤯
- Diversity helps!

Let's talk about

- The further synergy between informal and formal mathematics.
- How to apply AI in maths education?

References

[1] Lewkowycz, Aitor, et al. "Solving quantitative reasoning problems with language models." arXiv preprint arXiv:2206.14858 (2022).

[2] Zheng, Kunhao, Jesse Michael Han, and Stanislas Polu. "MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics." arXiv preprint arXiv:2109.00110 (2021).