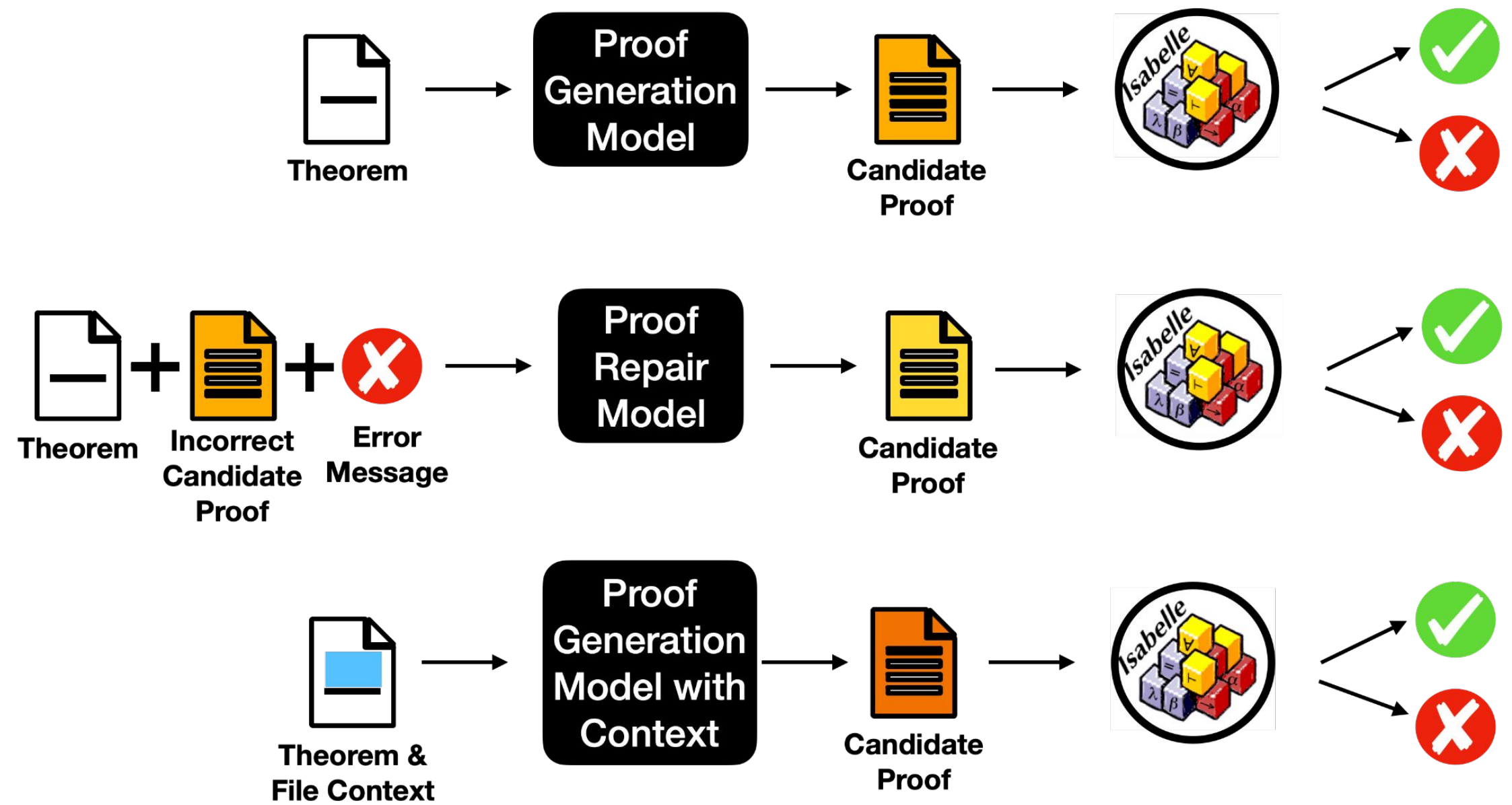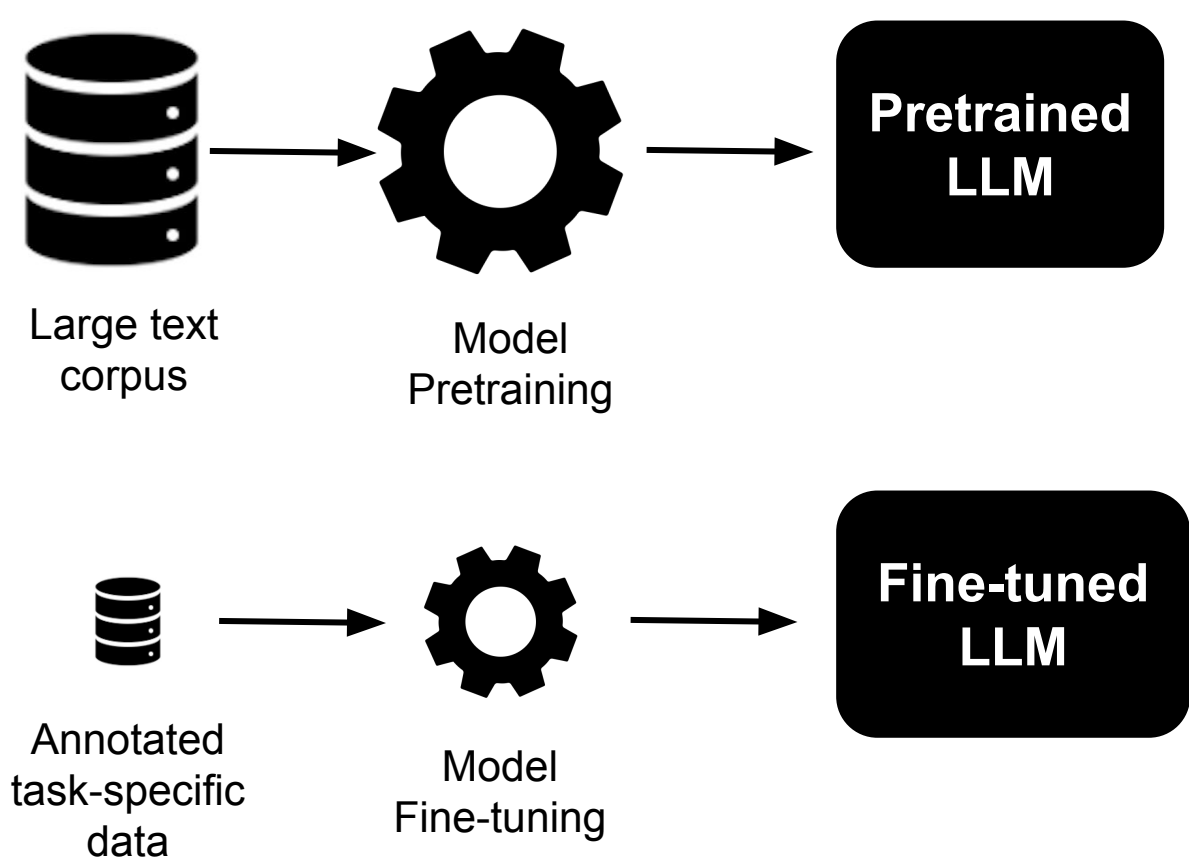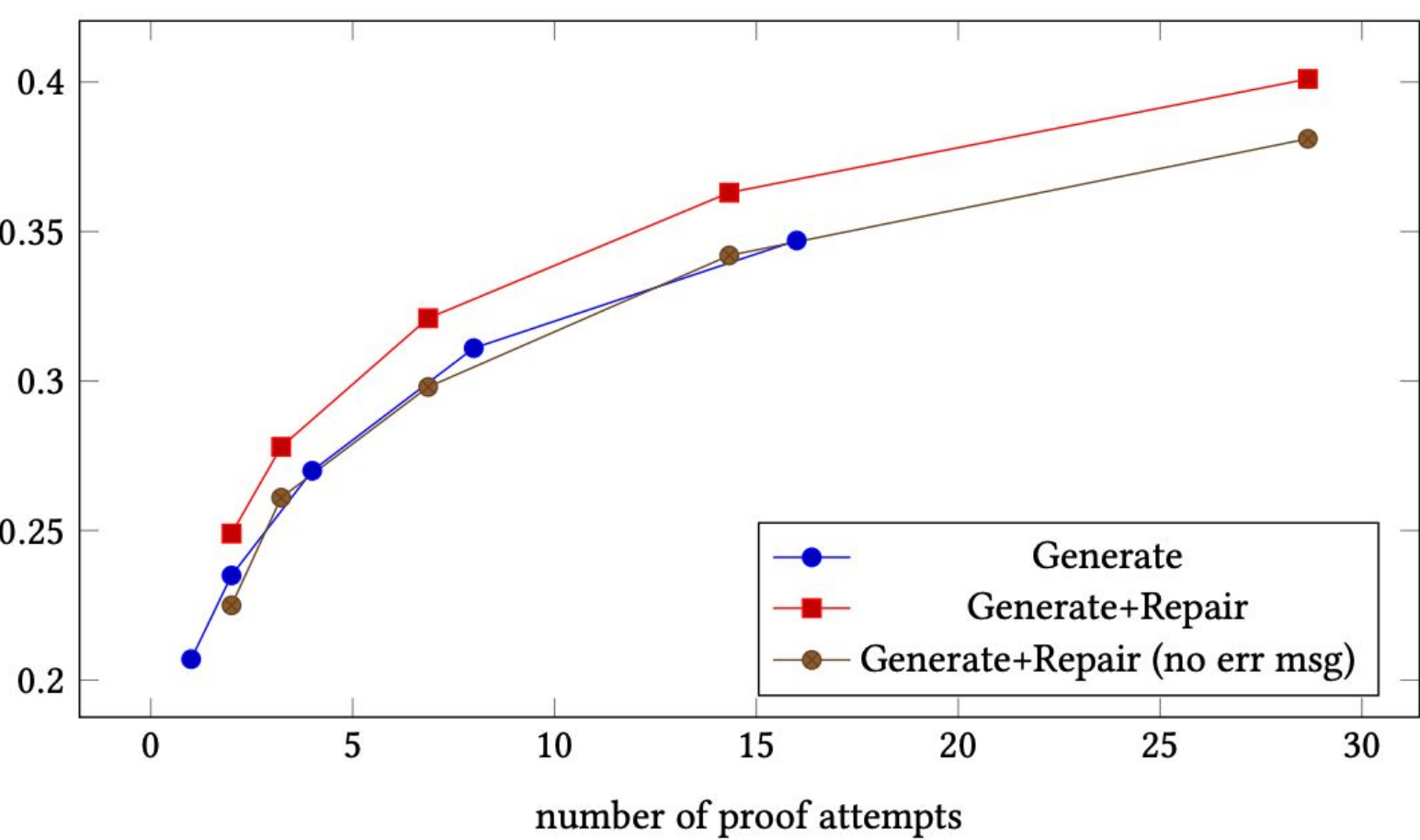# Fine-tuning and Prompting LLMs for Proof Synthesis and Repair

Emily First (UCSD), Talia Ringer (UIUC), Yuriy Brun (UMass Amherst)

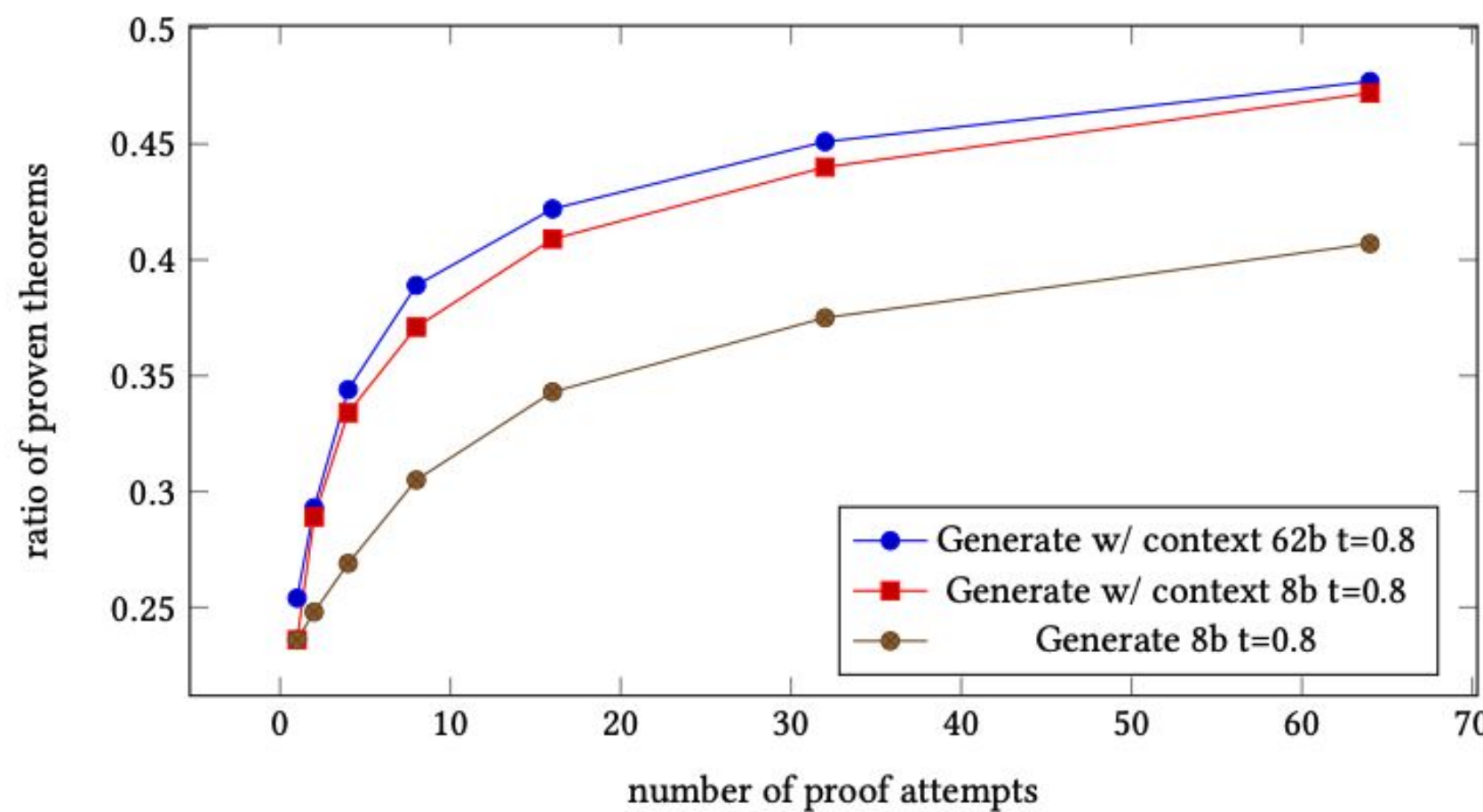## Baldur: Whole Proof Generation and Repair with LLMs

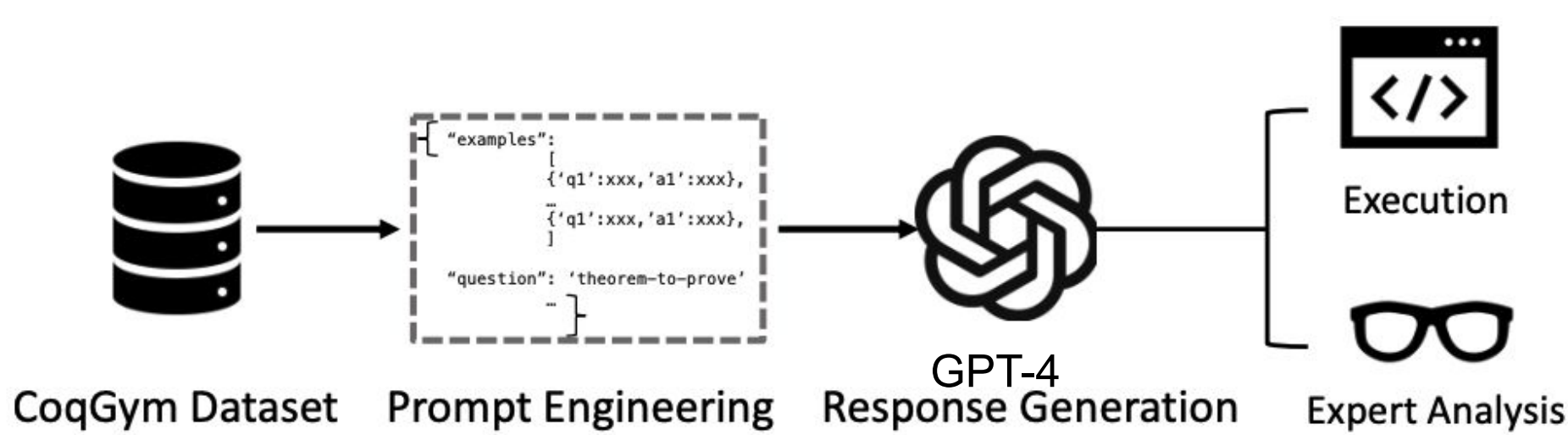Large text corpus → Model Pretraining → Pretrained LLM

Annotated task-specific data → Model Fine-tuning → Fine-tuned LLM

Theorem → Proof Generation Model → Candidate Proof → Isabelle → ✅ / ❌

Theorem + Incorrect Candidate Proof + Error Message → Proof Repair Model → Candidate Proof → Isabelle → ✅ / ❌

Theorem & File Context → Proof Generation Model with Context → Candidate Proof → Isabelle → ✅ / ❌

Generate vs Generate+Repair

ratio of proven theorems vs number of proof attempts

- Generate
- Generate+Repair
- Generate+Repair (no err msg)

Generate vs Generate w/ Context

ratio of proven theorems vs number of proof attempts

- Generate w/ context 62b t=0.8
- Generate w/ context 8b t=0.8
- Generate 8b t=0.8

Tool	PISA success rate
LLM with tree-search	39.0%
Baldur (8b w/ context)	47.5%
Thor (LLM+hammer with tree-search)	57.0%
Baldur + Thor	65.7%

Prompts for Proofs: Getting More out of LLMs for Proofs

CoqGym Dataset → Prompt Engineering → GPT-4 Response Generation → Execution / Expert Analysis

```
"examples":
[
    {'q1':xxx,'a1':xxx},
    {'q1':xxx,'a1':xxx},
]
"question": 'theorem-to-prove'
```

System Message

⚙ Generate proof of the theorem from user input in Coq. You should only generate proofs in Coq.

Few-shot examples from other projects
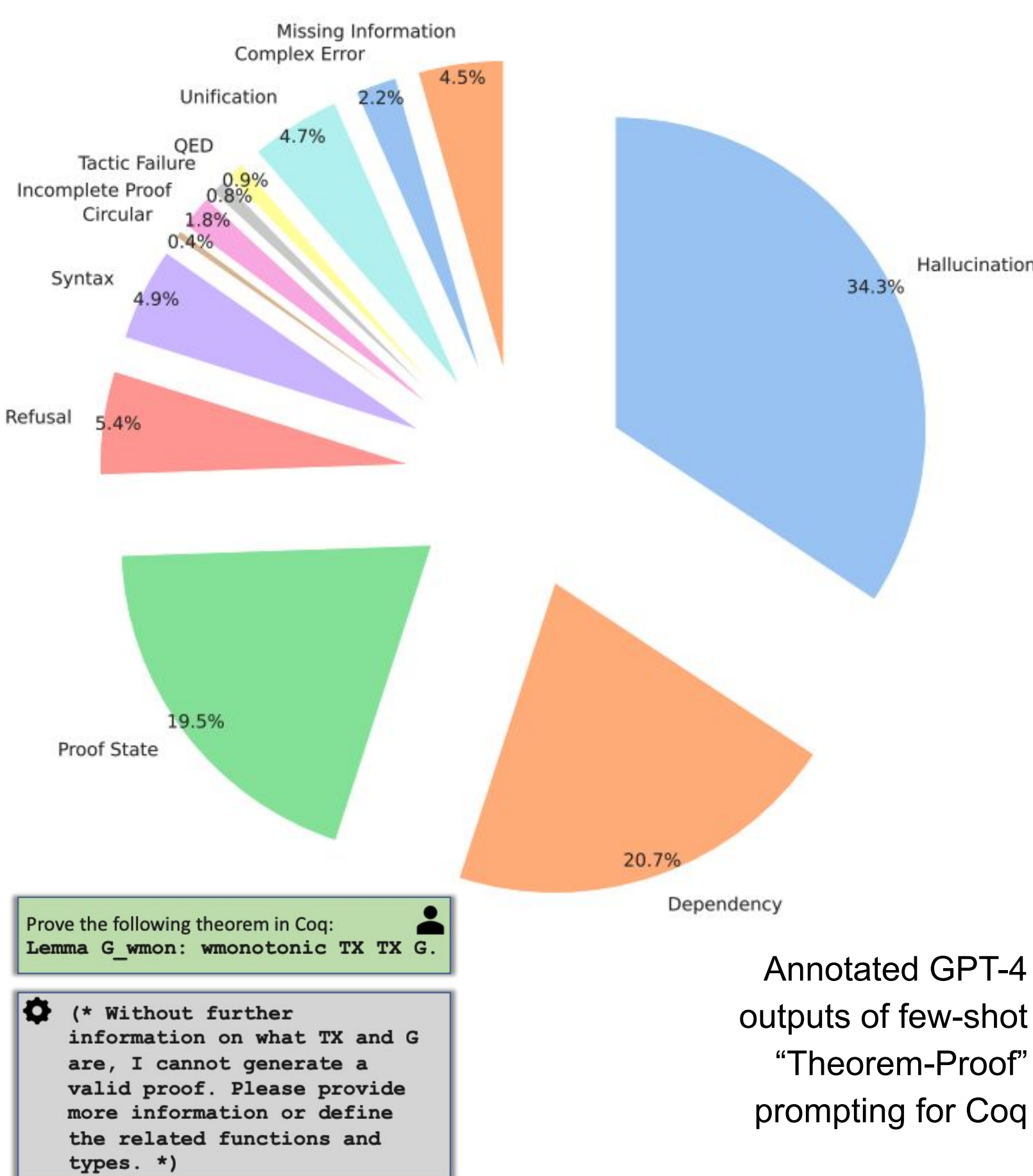
- Example Theorem 1 👤
- ⚙ Example Proof 1
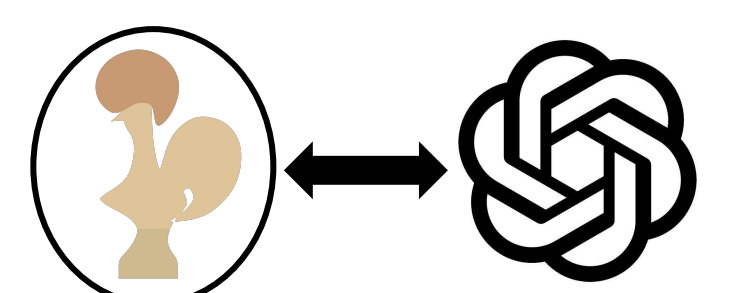- ......
- Example Theorem K 👤
- ⚙ Example Proof K

👤 Theorem in Test Project To Be Proven

Annotated GPT-4 outputs of few-shot "Theorem-Proof" prompting for Coq

Pie chart segments:
- Hallucination 34.3%
- Dependency 20.7%
- Proof State 19.5%
- Refusal 5.4%
- Syntax 4.9%
- Circular 0.4%
- Incomplete Proof 1.8%
- Tactic Failure 0.8%
- QED 0.9%
- Unification 4.7%
- Complex Error 2.2%
- Missing Information 4.5%

Prove the following theorem in Coq:
Lemma G_wmon: wmonotonic TX TX G.

⚙ (* Without further information on what TX and G are, I cannot generate a valid proof. Please provide more information or define the related functions and types. *)

New Direction (in-progress work)

System Message

Choose one of these actions to perform:
1. Edit the current code
2. Get definitions of any identifiers you don't know
3. Get the proof state at any point in the proof
4. Get the names of proven theorem/lemmas that are relevant to the identifier you specify

Facilitate conversation & search over action space