MiniF2F in Rocq

Automatic Translation Between Proof Assistants — A Case Study

Jules Viennot, Guillaume Baudart, Emilio Jesùs Gallego Arias, Marc Lelarge

IRIF, Université Paris Cité, Inria, CNRS
DI ENS, PSL University, Inria

2025-04-07

Introduction

Large Language Models and Reasoning

Use Large Language Models (LLMs) to generate reasoning in natural language or proof assistants (Rocq, Lean or Isabelle)

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$

DeepSeek-R1-Zero reasoning

```
Lemma ffx_eq_x_inj{A}:
  forall f:A->A,
  (forall x:A, f (f x) = x) ->
  forall x y:A, f x = f y -> x = y.
Proof.
  { intros f H1 x y H2.
    rewrite <- (H1 x).
    rewrite <- (H1 y).
    rewrite H2.
    reflexivity. }
Qed.
```

NLIR Rocq proof generation

Large Language Models and Reasoning

Use Large Language Models (LLMs) to generate reasoning in natural language or proof assistants (Rocq, Lean or Isabelle)

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

...

DeepSeek-R1-Zero reasoning

```
Lemma ffx_eq_x_inj{A}:
  forall f:A->A,
  (forall x:A, f (f x) = x) ->
  forall x y:A, f x = f y -> x = y.
Proof.
  { intros f H1 x y H2.
    rewrite <- (H1 x).
    rewrite <- (H1 y).
    rewrite H2.
    reflexivity. }
Qed.
```

NLIR Rocq proof generation

How to evaluate code generation methods?

Large Language Models and Reasoning

Use Large Language Models (LLMs) to generate reasoning in natural language or proof assistants (Rocq, Lean or Isabelle)

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\ldots$
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$
We started with the equation:
$\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
Next, I could square both sides again, treating the equation: $\cdots$
$\ldots$

DeepSeek-R1-Zero reasoning

```
Lemma ffx_eq_x_inj{A}:
  forall f:A->A,
  (forall x:A, f (f x) = x) ->
  forall x y:A, f x = f y -> x = y.
Proof.
  { intros f H1 x y H2.
    rewrite <- (H1 x).
    rewrite <- (H1 y).
    rewrite H2.
    reflexivity. }
Qed.
```

NLIR Rocq proof generation

How to evaluate code generation methods? $\Rightarrow$ benchmark datasets

MiniF2F

What is it?

Popular benchmark for ML based code generation in proof assistants

MiniF2F

What is it? Popular benchmark for ML based code generation in proof assistants

What is it made of?

488 exercises from olympiads (AMC, AIME, IMO) + high-school & undergraduate maths classes

mathd_numbertheory_227: Angela's problem


```
{
  "problem_name": "mathd_numbertheory_227",
  "informal_statement": "One morning each member of Angela's family drank an 8-ounce mixture of coffee with milk. The amounts of coffee and milk varied from cup to cup, but were never zero. Angela drank a quarter of the total amount of milk and a sixth of the total amount of coffee. How many people are in the family? Show that it is 5.",
  "informal_proof": "..."
}
```

What is it? Popular benchmark for ML based code generation in proof assistants

What is it made of? 488 high-school level maths exercises

What languages are supported?

Lean, Isabelle and Metamath ⇒ not Rocq

Angela's problem in Isabelle

```
theorem mathd_numbertheory_227:
  fixes x y n ::nat
  assumes "x / 4 + y / 6 = (x + y) / n"
    and "n\<noteq>0"
    and "x\<noteq>0"
    and "y\<noteq>0"
  shows "n = 5"
  sorry
end
```

Angela's problem in Lean

```
theorem mathd_numbertheory_227
  (x y n : ℕ+)
  (h₀ : ↑x / (4:ℝ) + y / 6 = (x + y) / n) :
  n = 5 :=
begin
  sorry
end
```

Our Goal

Informal description	+	Isabelle version	+	Lean version
...		...		...

Our Goal

Informal description
...

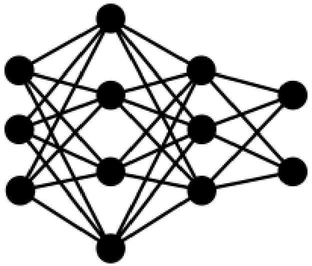
+

Isabelle version
...

+

Lean version
...

prompt



LLMs

Our Goal

Informal description
...

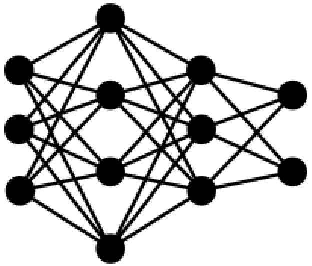
+

Isabelle version
...

+

Lean version
...

prompt



LLMs

=

Rocq version
...

Methodology

Models

Which models?

providers	models	open weights	chain of thought
OpenAI	GPT-4o mini	x	x
	o1-mini	x	o
	o1	x	o
Anthropic	Claude 3.5 Sonnet	x	x

GPT-4o mini $<$ Claude 3.5 Sonnet $<$ o1-mini $<$ o1

Models

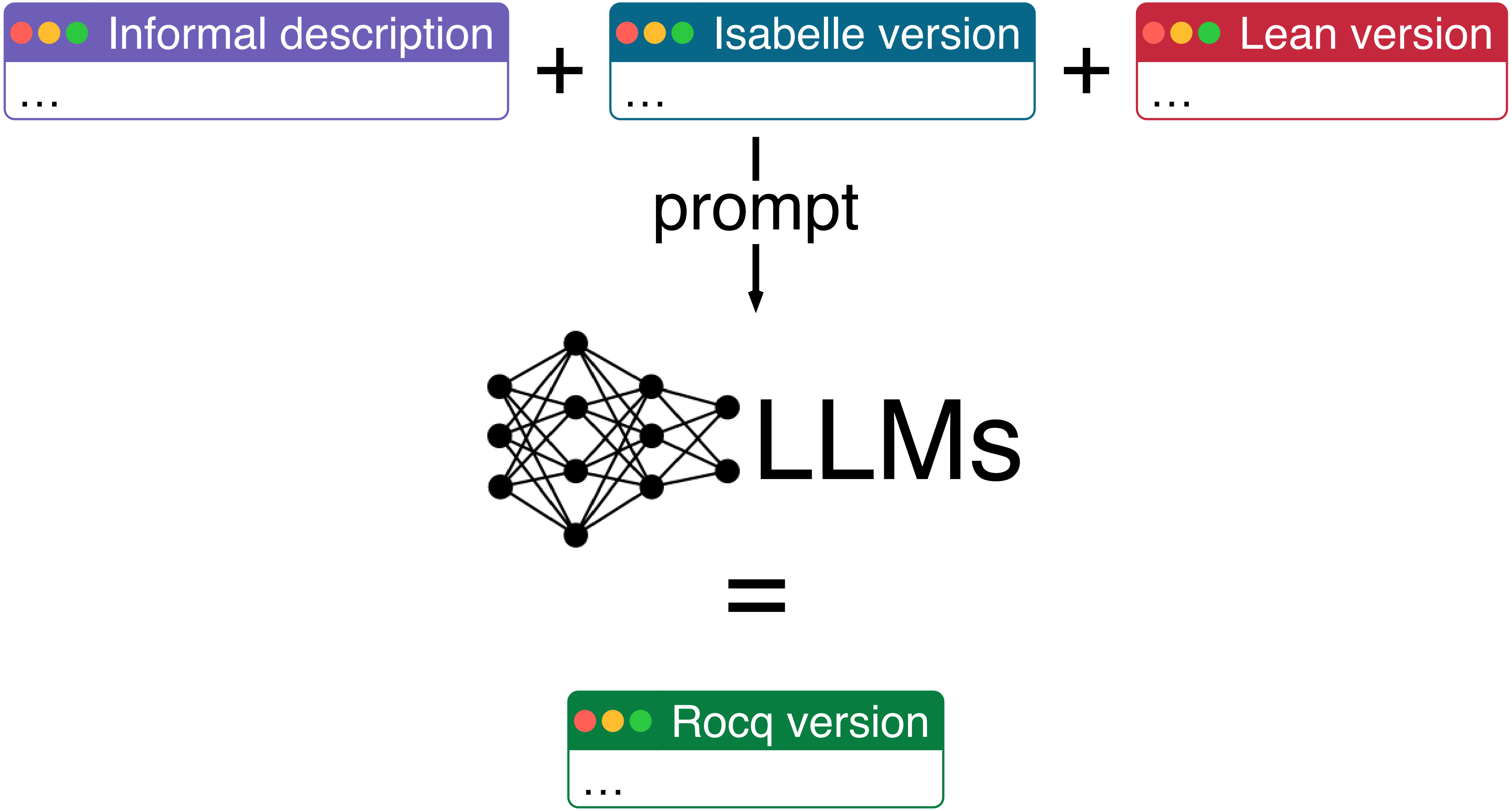
Which models?

providers	models	open weights	chain of thought
OpenAI	GPT-4o mini	x	x
	o1-mini	x	o
	o1	x	o
Anthropic	Claude 3.5 Sonnet	x	x

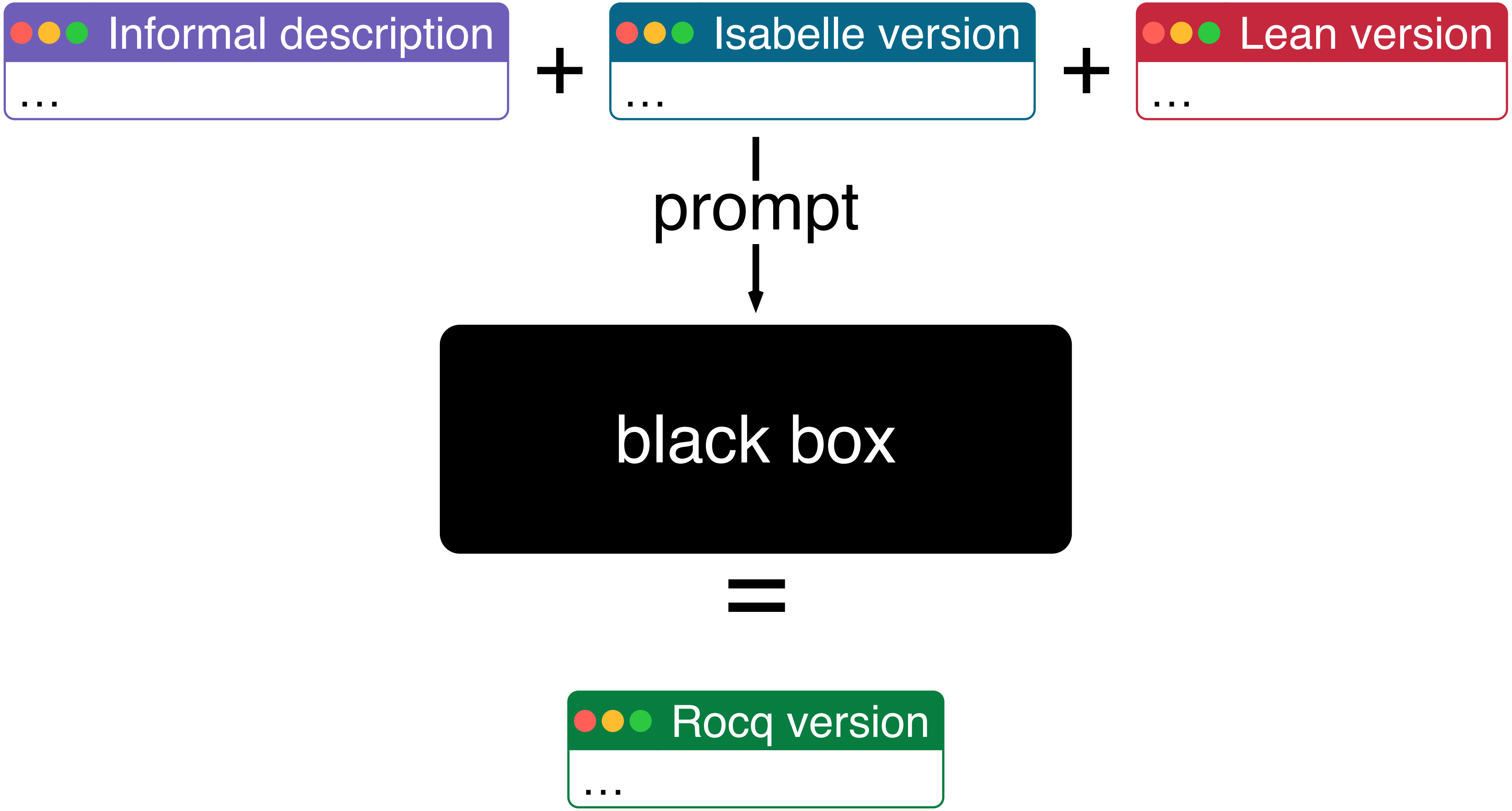
GPT-4o mini $<$ Claude 3.5 Sonnet $<$ o1-mini $<$ o1

No open weights models $\Rightarrow$ use them as **black boxes**

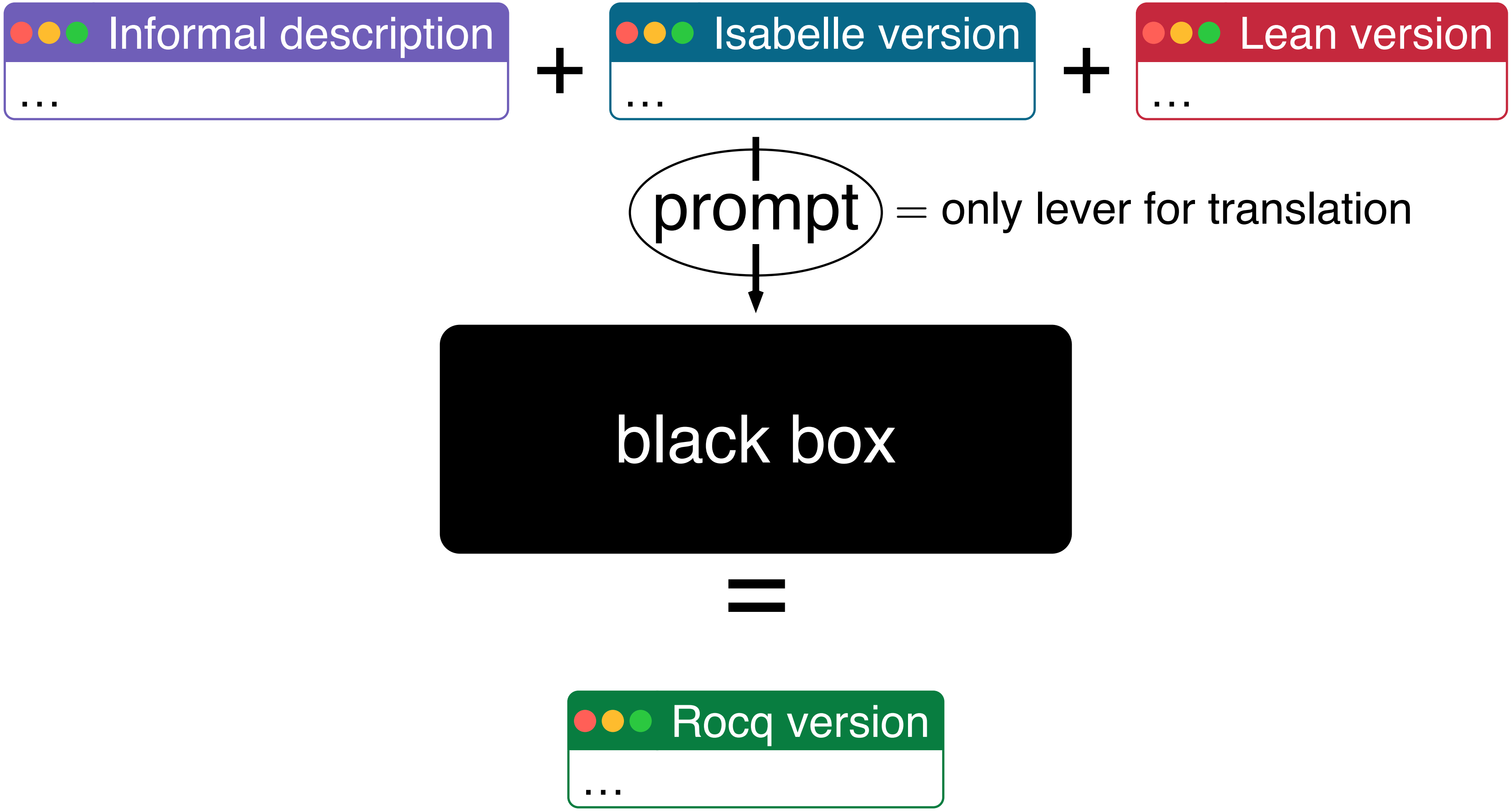
Strategy


Informal description
…
+
Isabelle version
…
+
Lean version
…
prompt
LLMs
=
Rocq version
…

Strategy


Informal description
...
+
Isabelle version
...
+
Lean version
...
prompt
black box
=
Rocq version
...

Strategy



Informal description
...
+
Isabelle version
...
+
Lean version
...
prompt
= only lever for translation
black box
=
Rocq version
...

3 Stages and human checking

3 stages, each stage is comprised of several steps

step: a model attempts to translate all theorems untranslated so far

3 Stages and human checking

3 stages, each stage is comprised of several steps

step: a model attempts to translate all theorems untranslated so far

At the end of each step, a **human** ensures all translated theorems are correct

Rocq example 1

```
Require Import Coq.Reals.Reals.
Require Import Coq.ZArith.ZArith.
Open Scope R_scope.
Open Scope Z_scope.

Parameter Rfloor : R -> Z.
Parameter big_sum : forall (m n : nat) (f : nat -> Z), Z.

Theorem aime_1991_p6 : forall (r : R),
    (big_sum 19 91 (fun k => Rfloor (r + (IZR (Z.of_nat k) / 100)))) = 546%Z ->
    Rfloor (100 * r) = 743%Z.
Proof.
Admitted.
```

3 Stages and human checking

3 stages, each stage is comprised of several steps

step: a model attempts to translate all theorems untranslated so far

At the end of each step, a **human** ensures all translated theorems are correct

Rocq example 1

```
Require Import Coq.Reals.Reals.
Require Import Coq.ZArith.ZArith.
Open Scope R_scope.
Open Scope Z_scope.

Parameter Rfloor : R -> Z.
Parameter big_sum : forall (m n : nat) (f : nat -> Z), Z.

Theorem aime_1991_p6 : forall (r : R),
    (big_sum 19 91 (fun k => Rfloor (r + (IZR (Z.of_nat k) / 100)))) = 546%Z ->
    Rfloor (100 * r) = 743%Z.
Proof.
Admitted.
```

3 Stages and human checking

3 stages, each stage is comprised of several steps

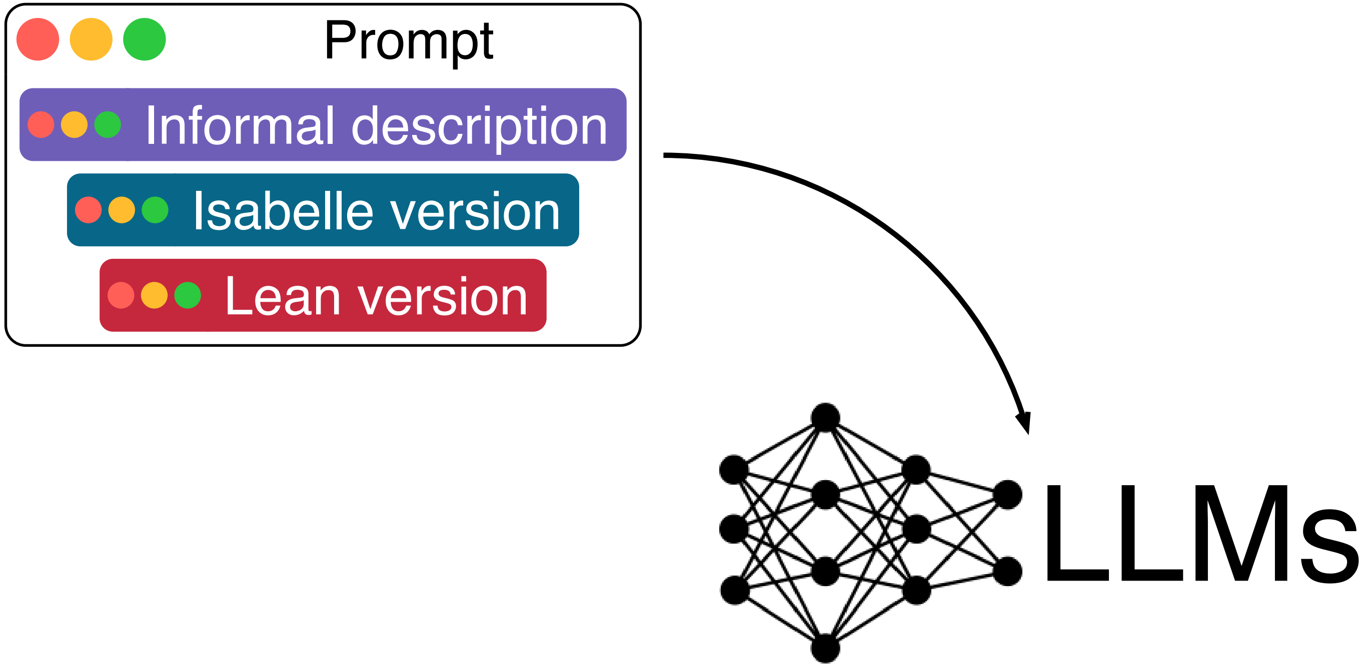
step: a model attempts to translate all theorems untranslated so far

At the end of each step, a **human** ensures all translated theorems are correct

If a theorem is considered incorrect, it is **put back** with the untranslated theorems

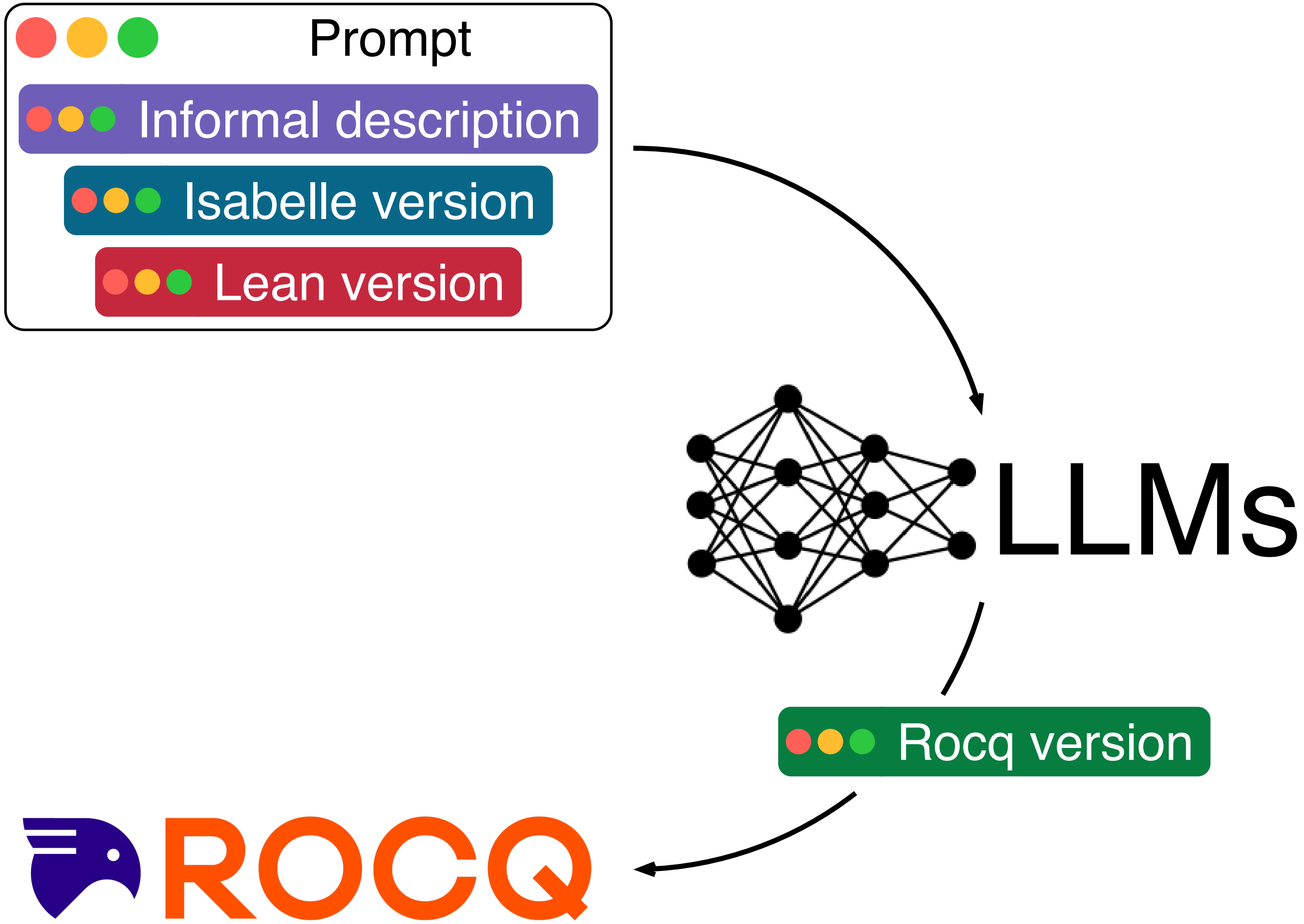
Stage 1: one-shot prompting

Pipeline

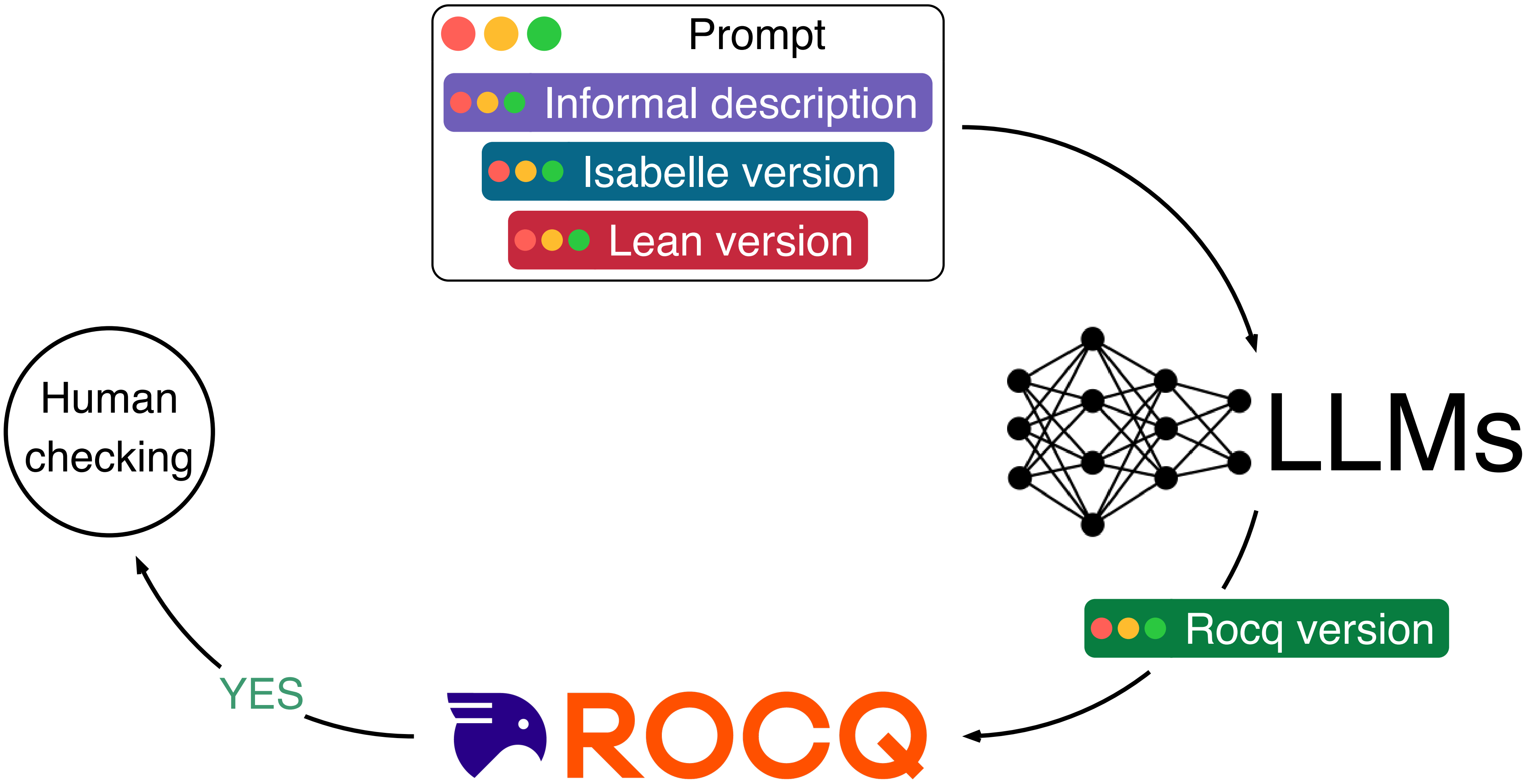

Prompt
Informal description
Isabelle version
Lean version
LLMs

Pipeline

Stage 1: one-shot prompting



Pipeline


Prompt
Informal description
Isabelle version
Lean version
LLMs
Rocq version
ROCQ
YES
Human checking

Example

Informal description

```json
{
  "problem_name": "mathd_numbertheory_227",
  "informal_statement": "One morning each member of Angela's family drank an 8-ounce mixture of coffee with milk. The amounts of coffee and milk varied from cup to cup, but were never zero. Angela drank a quarter of the total amount of milk and a sixth of the total amount of coffee. How many people are in the family? Show that it is 5.",
  "informal_proof": "..."
}
```

Isabelle version

```
theorem mathd_numbertheory_227:
  fixes x y n ::nat
  assumes "x / 4 + y / 6 = (x + y) / n"
    and "n\<noteq>0"
    and "x\<noteq>0"
    and "y\<noteq>0"
  shows "n = 5"
  sorry
end
```

Lean version

```
theorem mathd_numbertheory_227
  (x y n : ℕ+)
  (h₀ : ↑x / (4:ℝ) + y / 6 = (x + y) / n) :
  n = 5 :=
begin
  sorry
end
```

Example

Informal description

...

Isabelle version

...

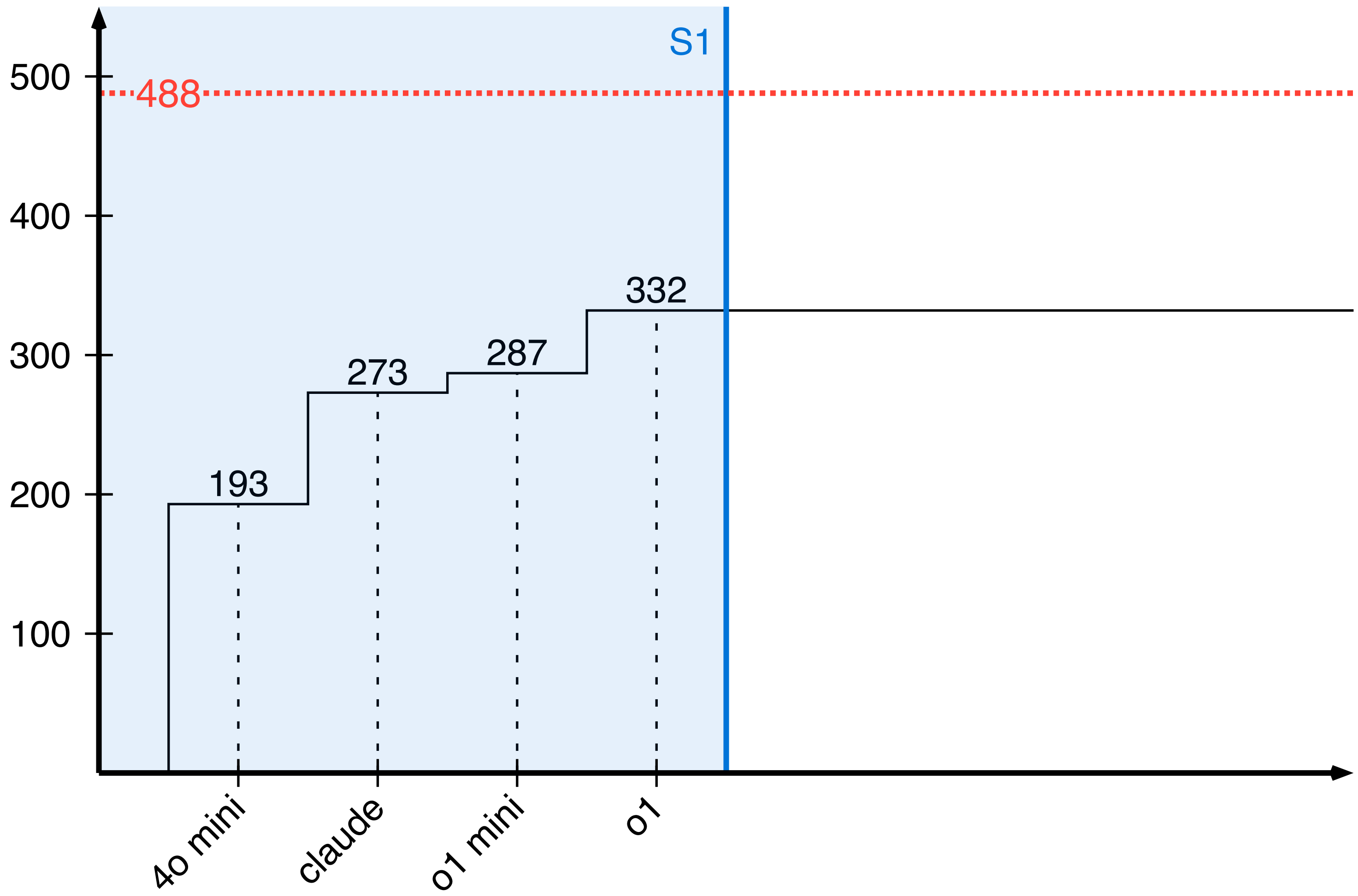
Lean version

...

Rocq version

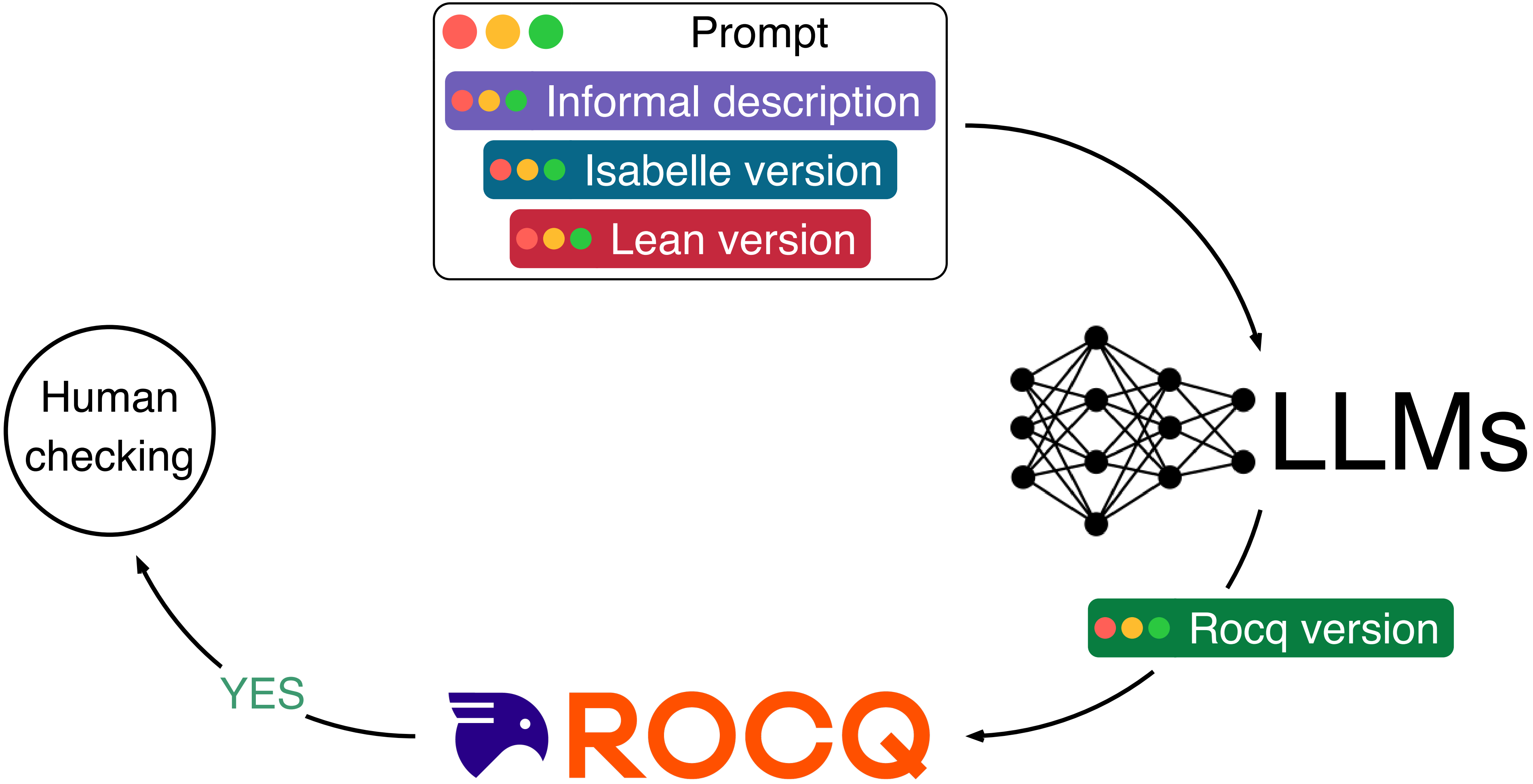
```
Require Import Coq.Reals.Reals.

Open Scope R_scope.

Theorem mathd_numbertheory_227:
  forall (x y n : nat), x <> 0%nat -> y <> 0%nat -> n <> 0%nat ->
    (INR x / 4 + INR y / 6 = (INR x + INR y) / INR n) -> n = 5%nat.
Proof.
Admitted.
```

Results

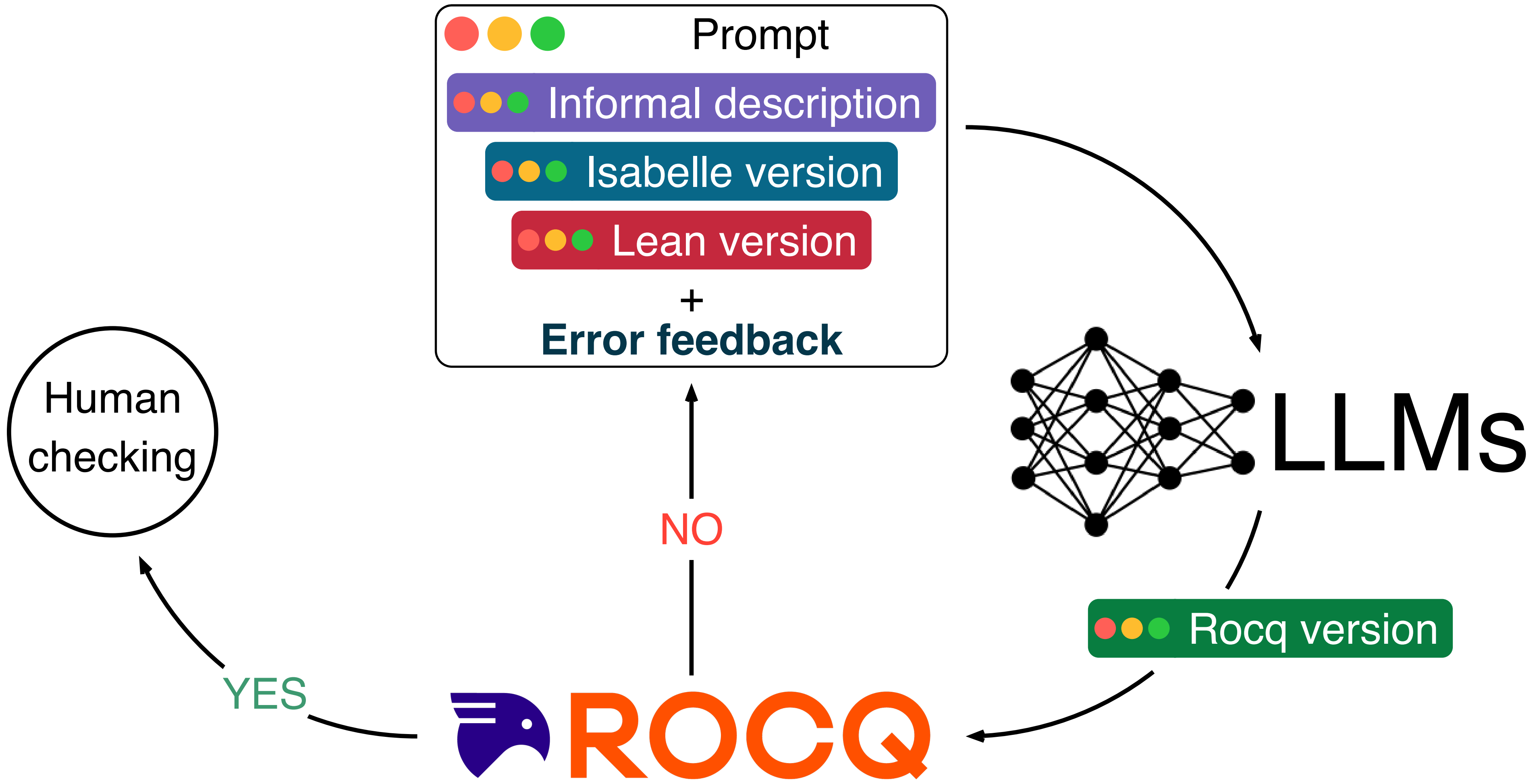

S1
488
500
400
300
200
100
193
273
287
332
4o mini
claude
o1 mini
o1

Stage 2: multi-turn with errors

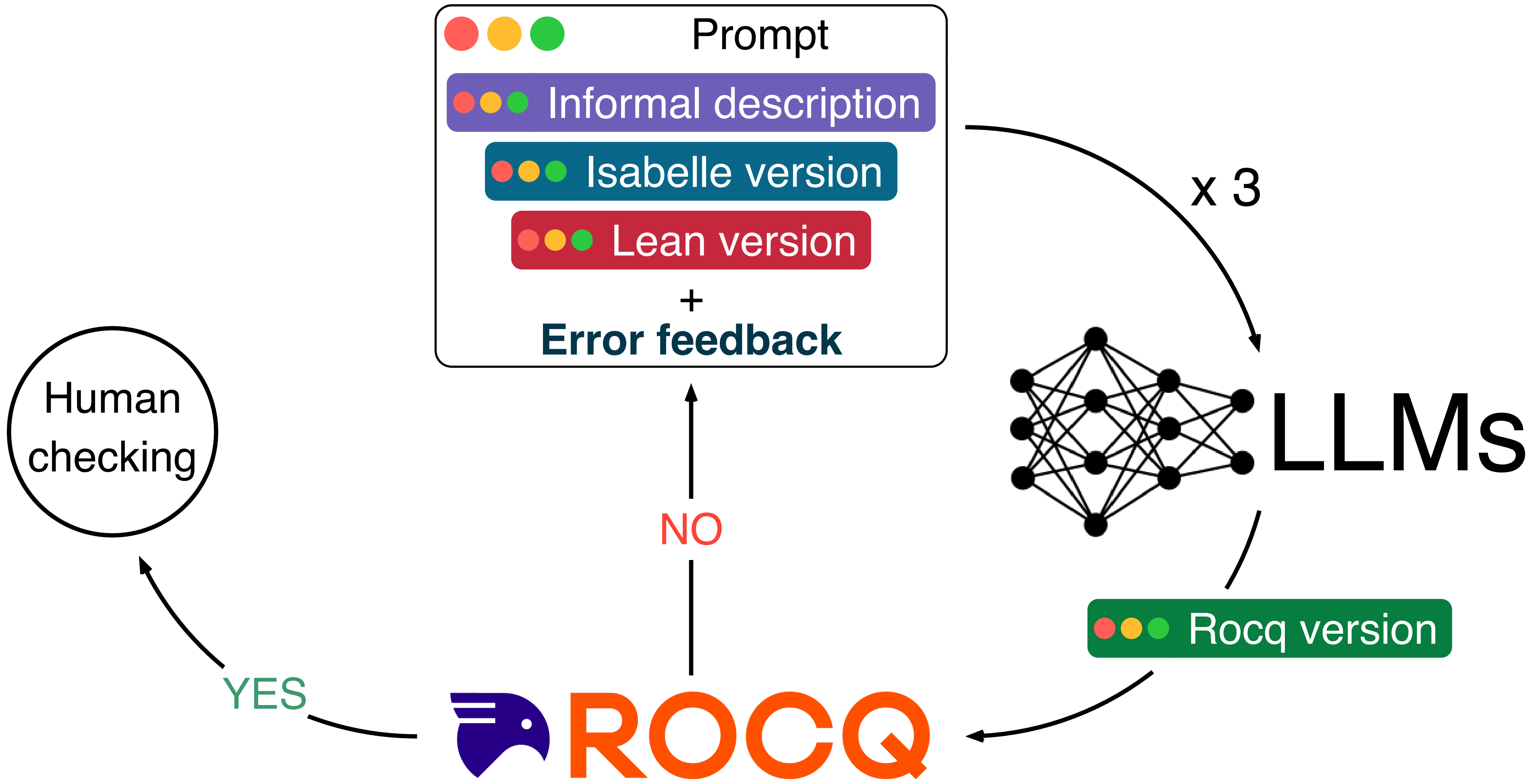
Pipeline


Prompt
Informal description
Isabelle version
Lean version
LLMs
Rocq version
ROCQ
YES
Human checking

Pipeline


Prompt
Informal description
Isabelle version
Lean version
+
Error feedback
LLMs
Rocq version
ROCQ
NO
YES
Human checking

Pipeline


Prompt
Informal description
Isabelle version
Lean version
+
Error feedback
x 3
LLMs
Rocq version
ROCQ
NO
YES
Human checking

Example

Rocq example 2 : first attempt unsuccessful

```
Require Import Arith.

Theorem imo_1964_p1_1 :
  forall n : nat,
    (7 | (2^n - 1)%nat) -> (3 | n).
Proof.
Admitted.
```

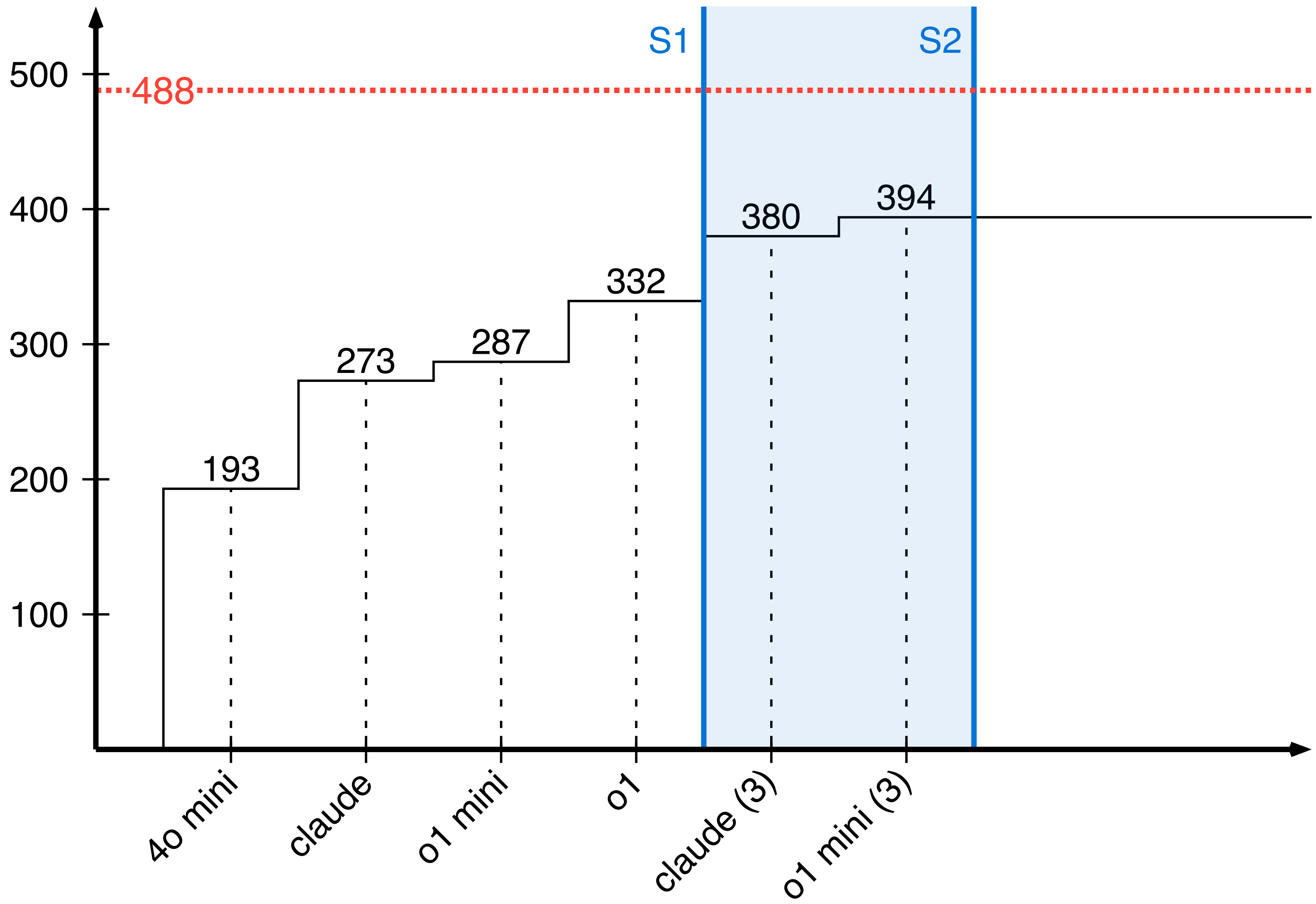
Errors :

Syntax error: ',' or ')' expected after [term level 200] (in [term]).

Rocq example 2 : second attempt successful

```
Require Import Arith.

Theorem imo_1964_p1_1 :
  forall n : nat,
    (Nat.divide 7 (2^n - 1)) -> (Nat.divide 3 n).
Proof.
Admitted.
```

Results


S1
S2
500
400
300
200
100
488
193
273
287
332
380
394
4o mini
claude
o1 mini
o1
claude (3)
o1 mini (3)

Stage 3: refined prompt

Introduction

Rocq example 3 : before stage 3

```
Require Import Coq.Complex.Reals.

Theorem mathd_algebra_302 :
  (Caux.I / 2)^2 = -(1 / 4).
Proof.
Admitted.
```

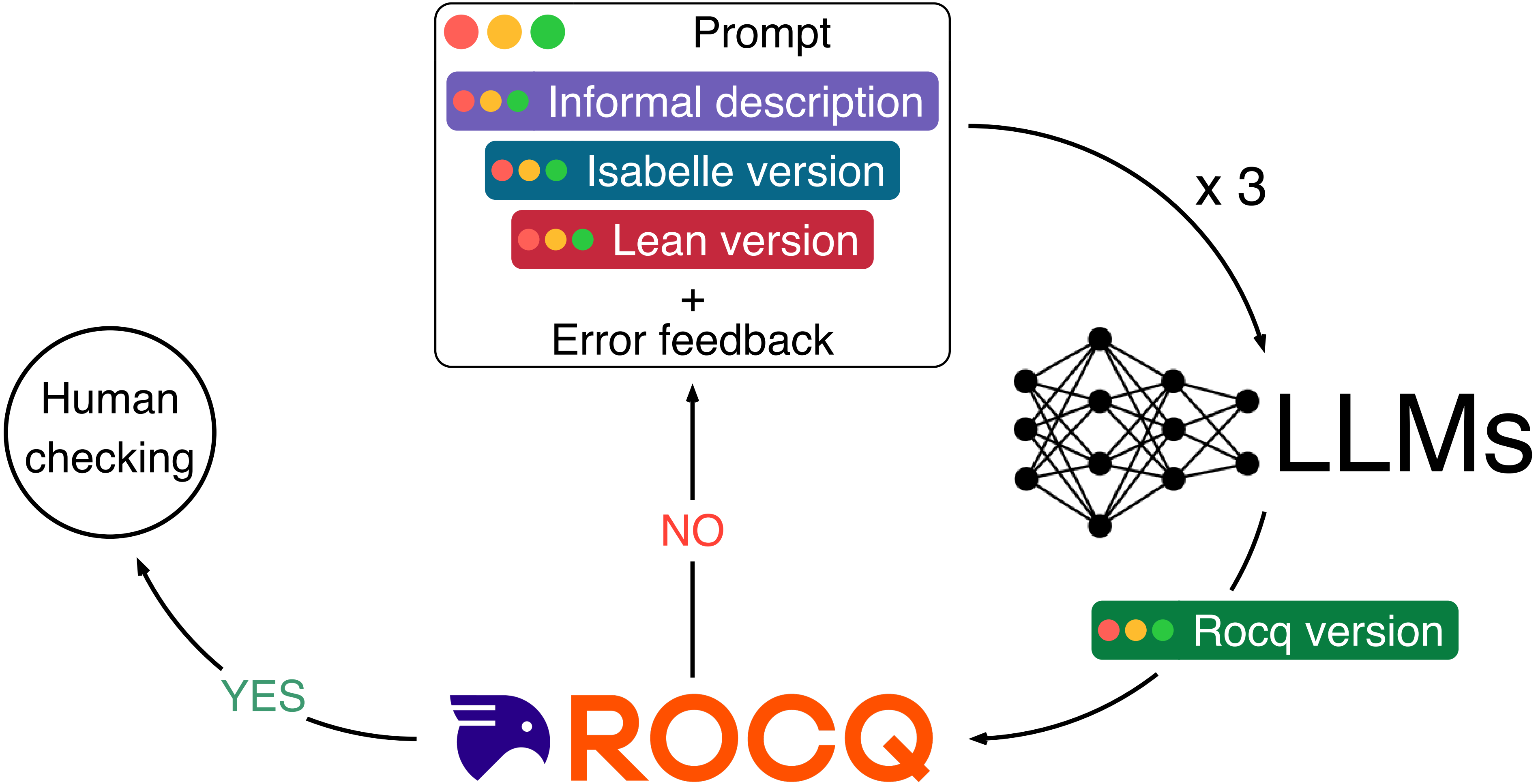
```
Errors :
Cannot find a physical path bound to
logical path Stdlib.Complex.Reals.
```

Rocq example 4 : before stage 3

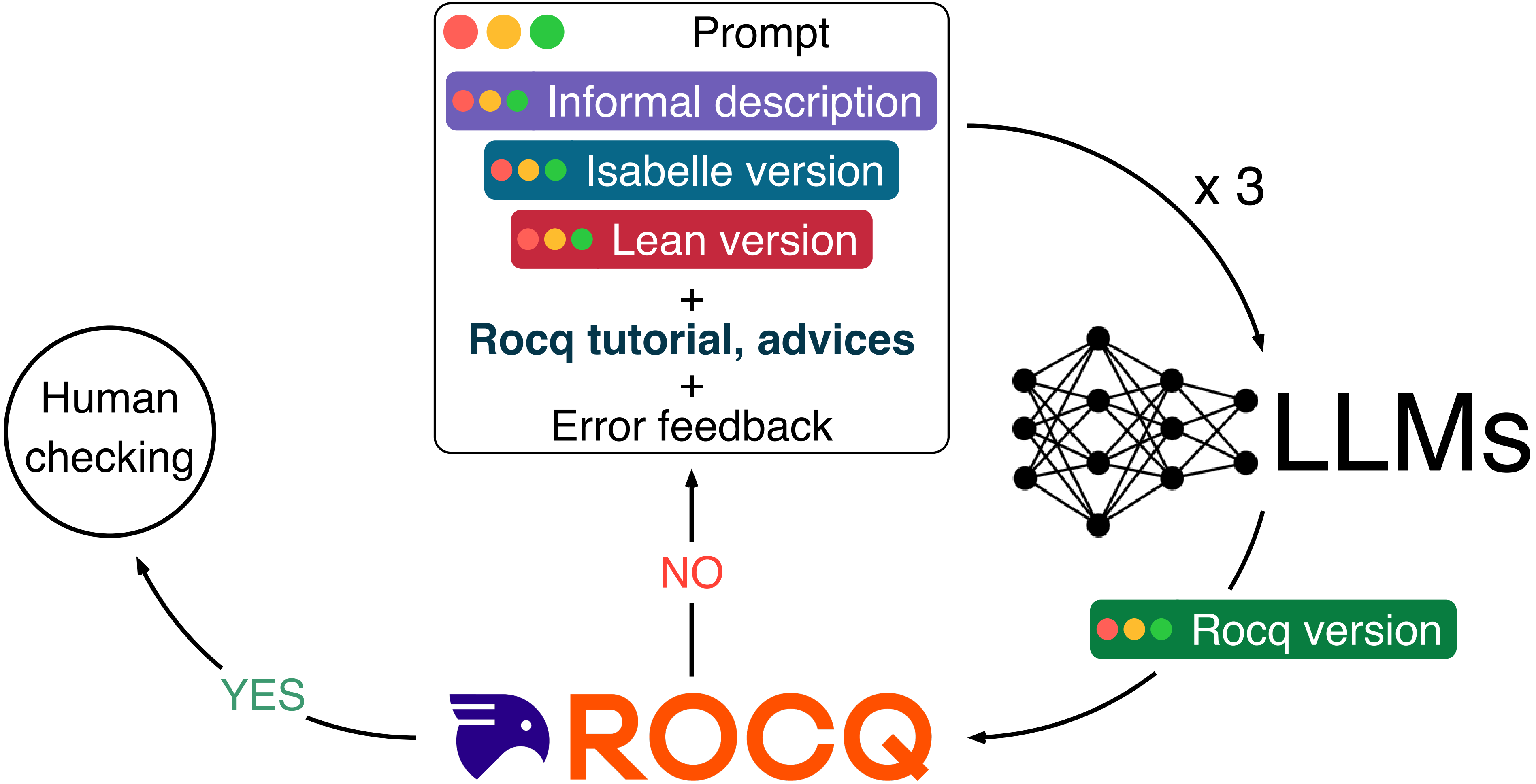
```
Require Import Reals.
Require Import Coquelicot.Coquelicot.
Require Import QArith.
Require Import ZArith.
Require Import List.

Open Scope R_scope.

Theorem amc12a_2020_p25 :
  forall (a : Q) (S : list R),
  (forall x : R, In x S <->
    (IZR (floor x) * (x - IZR (floor x))
      = (Q2R a) * (x ^ 2)))
    -> NoDup S
    -> fold_left Rplus S 0 = 420
    -> (Qnum a + Qden a)%Z = 929.
Proof.
Admitted.
```

```
Errors :
In environment
a : Q
S : list R
x : R
The term "x" has type "R" while it is
expected to have type "positive".
```

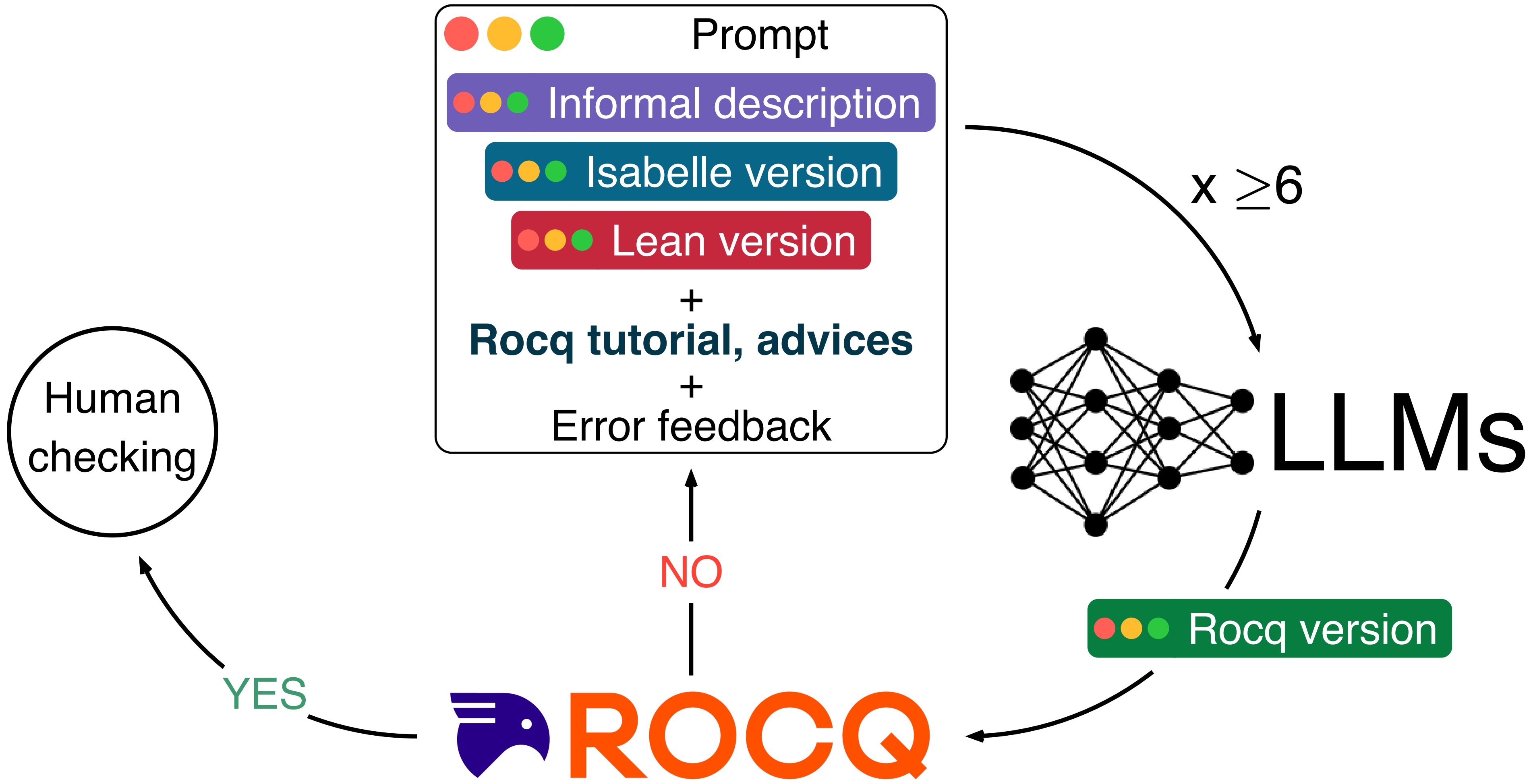
Pipeline


Prompt
Informal description
Isabelle version
Lean version
+
Error feedback
x 3
LLMs
Rocq version
ROCQ
NO
YES
Human checking

Pipeline


Prompt
Informal description
Isabelle version
Lean version
+
Rocq tutorial, advices
+
Error feedback
x 3
LLMs
Rocq version
ROCQ
NO
YES
Human checking

Pipeline


Prompt
Informal description
Isabelle version
Lean version
+
Rocq tutorial, advices
+
Error feedback
x ≥6
LLMs
Rocq version
NO
YES
Human checking
ROCQ

Examples

Rocq example 3 : before stage 3

```
Require Import Coq.Complex.Reals.



Theorem mathd_algebra_302 :
  (Caux.I / 2)^2 = -(1 / 4).
Proof.
Admitted.
```

Errors :

Cannot find a physical path bound to logical path Stdlib.Complex.Reals.

Examples

Rocq example 3 : after stage 3

```
Require Import Reals.
Require Import Coquelicot.Coquelicot.

Open Scope C_scope.

Theorem mathd_algebra_302 :
  (Ci / 2)^2 = - (1 / 4).
Proof.
Admitted.
```

Examples

Rocq example 4 : before stage 3

```
Require Import Reals.
Require Import Coquelicot.Coquelicot.
Require Import QArith.
Require Import ZArith.
Require Import List.
Open Scope R_scope.

Theorem amc12a_2020_p25 :
  forall (a : Q),
    forall (S : list R),
      (forall x : R, In x S <->
        (IZR (floor x) * (x - IZR (floor x))
          = (Q2R a) * (x ^ 2)))
    -> NoDup S
    -> fold_left Rplus S 0 = 420
    -> (Qnum a + Qden a)%Z = 929.
Proof.
Admitted.
```

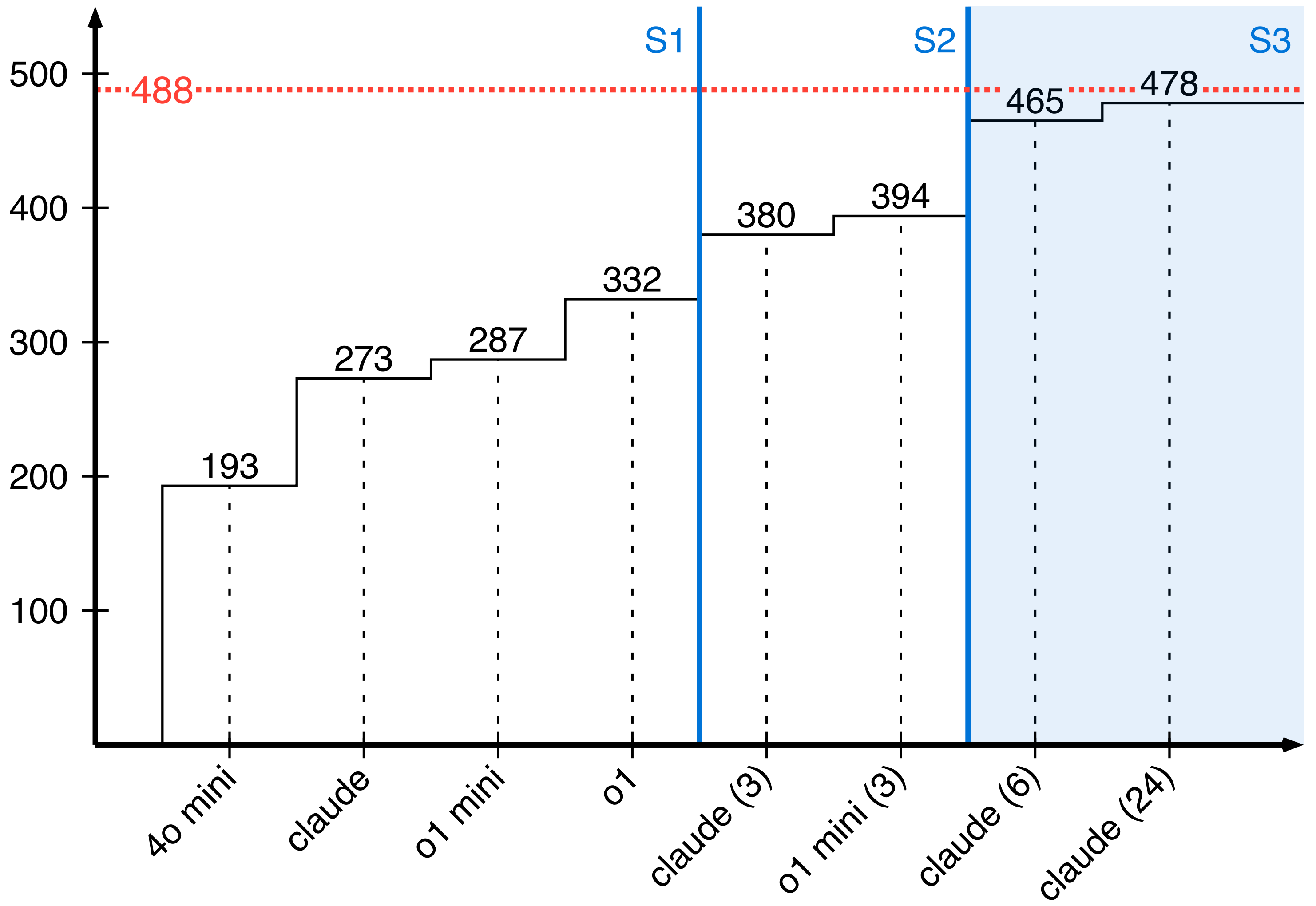
```
Errors :
In environment
a : Q
S : list R
x : R
The term "x" has type "R" while it is
expected to have type "positive".
```

Examples

Rocq example 4 : after stage 3

```
Require Import Reals.
Require Import Coquelicot.Coquelicot.
Require Import QArith.
Require Import ZArith.
Require Import List.
Open Scope R_scope.

Theorem amc12a_2020_p25 :
  forall (a : Q),
    forall (S : list R),
      (forall x : R, In x S <->
        (IZR (Int_part x) * (x - IZR (Int_part x))
          = Q2R a * Rpower x 2))
    -> NoDup S
    -> fold_left Rplus S 0 = 420
    -> (Z.pos (Qden a) + Qnum a = 929)%Z.
Proof.
Admitted.
```

Results



S1
S2
S3
488
500
400
300
200
100
193
273
287
332
380
394
465
478
4o mini
claude
o1 mini
o1
claude (3)
o1 mini (3)
claude (6)
claude (24)

Evaluation

Questions

RQ1 Does a better model really performs better?

RQ2 Does changing the amount of information on a theorem changes the performance of the model?

RQ3 Does the translated statement of the theorem make the proof harder to write?

RQ1 - Models comparison

RQ1: Does a better model really performs better?
$\rightarrow$ **GPT-4o mini** vs **o1-mini**

Select 100 theorems, 50 of which GPT-4o mini translated at stage 1

RQ1 - Models comparison

RQ1: Does a better model really performs better?
$\rightarrow$ **GPT-4o mini** vs **o1-mini**

Comparison: **pass@1** $=$ one one-shot prompting (stage 1) on the 100 theorems

	o1-mini success	o1-mini fail	Total
GPT-4o mini success	28	22	50
GPT-4o mini fail	10	40	50
Total	38	62	100

$\Rightarrow$ GPT-4o mini $>$ o1-mini?

RQ1 - Models comparison

RQ1: Does a better model really performs better?
$\rightarrow$ **GPT-4o mini** vs **o1-mini**

Comparison: **pass@3** $=$ three one-shot prompting (stage 1) on the 100 theorems

	o1-mini success	o1-mini fail	Total
GPT-4o mini success	**58**	7	65
GPT-4o mini fail	6	**29**	35
Total	64	36	100

$\Rightarrow$ GPT-4o mini $\approx$ o1-mini

$\Rightarrow$ notion of **easy** and **hard** translations

RQ2 - Ablation study

RQ2: Does changing the amount of information on a theorem changes the performance of the model?

Comparison: **one** and **three** one-shot prompting on the 100 theorems with o1-mini

RQ2 - Ablation study

RQ2: Does changing the amount of information on a theorem changes the performance of the model?

Comparison: **one** and **three** one-shot prompting on the 100 theorems with o1-mini

Information in the prompt	Pass@1	Pass@3
informal description + isabelle version + lean version	38%	64%
informal description	51%	75%
isabelle version + lean version	41%	62%
lean version	42%	60%

$\Rightarrow$ only informal description $>$ rest

$\Rightarrow$ all information $\approx$ only code versions $\approx$ only lean version

RQ3 - Audit: introduction

RQ3: Does the translated statement of the theorem make the proof harder to write?

Method: ask Rocq users to review batch of 25 translated theorems

RQ3 - Audit: introduction

RQ3: Does the translated statement of the theorem make the proof harder to write?

$\Rightarrow$ hard question: what does it really mean?

Method: ask Rocq users to review batch of 25 translated theorems

RQ3 - Audit: introduction

RQ3: Does the translated statement of the theorem make the proof harder to write?

⇒ hard question: what does it really mean?

Method: ask Rocq users to review batch of 25 translated theorems

⇒ relying on their judgement: from **no** badly written theorems for some to **half** the badly written theorems for others

RQ3 - Audit: other goals

- Finding unnoticed errors

Rocq example 4 : before the audit

```
Require Import Reals.
Require Import Coquelicot.Coquelicot.
Require Import QArith.
Require Import ZArith.
Require Import List.
Open Scope R_scope.

Theorem amc12a_2020_p25 :
  forall (a : Q),
    forall (S : list R),
      (forall x : R, In x S <->
        (IZR (Int_part x) * (x - IZR (Int_part x)) = Q2R a * Rpower x 2))
    -> NoDup S
    -> fold_left Rplus S 0 = 420
    -> (Z.pos (Qden a) + Qnum a = 929)%Z.
Proof.
Admitted.
```

RQ3 - Audit: other goals

- Finding unnoticed errors

Rocq example 4 : after the audit

```
Require Import Reals.



Require Import List.
Open Scope R_scope.

Theorem amc12a_2020_p25 :
  forall (p q : nat), Nat.gcd p q = 1%nat ->
    forall (S : list R),
      (forall x : R, In x S <->
        (IZR (Int_part x) * (x - IZR (Int_part x)) = INR p / INR q * Rpower x 2))
    -> NoDup S
    -> fold_left Rplus S 0 = 420
    -> (p + q = 929)%nat.
Proof.
Admitted.
```

RQ3 - Audit: other goals

- Finding unnoticed errors

- Better alignment with the informal description

Informal description

```
{ "informal_statement": "What is the maximum value of (2^t - 3t) * t / 4^t  for real values of t? Show that it is 1 / 12." }
```

Rocq example 6 : before the audit

```
Require Import Coq.Reals.Reals.
Open Scope R_scope.

Theorem amc12b_2020_p22 : forall t : R,
  ((exp (t * ln 2) - 3 * t) * t) / (exp (t * ln 4)) <= 1 / 12.

Proof.
Admitted.
```

RQ3 - Audit: other goals

- Finding unnoticed errors
- Better alignment with the informal description

Informal description

{ "informal_statement": "What is the maximum value of (2^t - 3t) * t / 4^t for real values of t? Show that it is 1 / 12." }

Rocq example 6 : after the audit

```
Require Import Coq.Reals.Reals.
Open Scope R_scope.

Theorem amc12b_2020_p22 : forall t : R,
  ((exp (t * ln 2) - 3 * t) * t) / (exp (t * ln 4)) <= 1 / 12 /\
  exists t,  ((exp (t * ln 2) - 3 * t) * t) / (exp (t * ln 4)) = 1 / 12.
Proof.
Admitted.
```

RQ3 - Audit: other goals

- Finding unnoticed errors
- Better alignment with the informal description
- Removing useless content or write better syntax (e.g. currying)

Rocq example 7 : before the audit

```
Require Import PeanoNat.

Theorem aime_1991_p1 :
  forall (x y : nat), (0 < x)%nat -> (0 < y)%nat ->
    (x * y + x + y = 71) ->
    (x^2 * y + x * y^2 = 880) ->
    (x^2 + y^2 = 146).
Proof.
Admitted.
```

RQ3 - Audit: other goals

- Finding unnoticed errors
- Better alignment with the informal description
- Removing useless content or write better syntax (e.g. currying)

Rocq example 7 : after the audit

```
Theorem aime_1991_p1 :
  forall (x y : nat),
    (x * y + x + y = 71) ->
    (x*x * y + x * y*y = 880) ->
    (x*x + y*y = 146).
Proof.
Admitted.
```

RQ3 - Audit: results

Audit so far: **150** problems $\approx$ **31%** of the dataset

Results so far:

Answers	**Percentages**
Error	2%
Reformulation	4%
Syntax	17.3%
Valid	76.7%
Proof	18.7%

Conclusion

Main Lessons

- **Feedback importance**:
  - big improvement by adding previous failed attempts
  - final errors are often due to the incapacity to correctly use previous attempts

Rocq example 5 : unproven

```
Require Import Reals.

Theorem aime_1988_p8 :
  forall (f : nat -> nat -> R),
  (forall x, (0 < x)%nat -> f x x = INR x) ->
  (forall x y, (0 < x)%nat /\ (0 < y)%nat -> f x y = f y x) ->

  (forall x y, (0 < x)%nat /\ (0 < y)%nat ->
    (INR (Nat.add x y)) * (f x y) = (INR y) * (f x (Nat.add x y))) ->

  f 14 52 = INR 364.
Proof.
Admitted.
```

```
Errors :
In environment
f : nat -> nat -> R
x : nat
y : nat
The term "INR (x + y)"
has type "R" while it
is expected to have
type "nat".
```

Main Lessons

- **Feedback importance**:
  - big improvement by adding previous failed attempts
  - final errors are often due to the incapacity to correctly use previous attempts

- **Indications importance**:
  - the fewer the examples on internet, the worst the LLMs
    $\rightarrow$ scraping on github:

Domains	Number of files
nat	$\sim$ 80k
reals	$\sim$ 8k
complex numbers	$\sim$ 500

  - LLMs struggle with types and scopes

Thank you!

Help us by **participating** to the theorems audit!

Contact us at llm4coq@gmail.com

The dataset is available at https://github.com/LLM4Rocq/miniF2F-rocq and on HuggingFace 🤗 at https://huggingface.co/datasets/LLM4Rocq/miniF2F-rocq