LLM4Docq: Bootstrapping Documentation for MathComp

Théo STOSKOPF, Jules VIENNOT, Cyril COHEN

Inria

UNIVERSITÉ DE LYON

ENS DE LYON

Context

MathComp, a library of formalized mathematics in Rocq, doesn't have any docstrings.

Yet, docstrings can be very useful for:

- learning a library
- contributing to it
- building dataset for deep learning.

Annotating each element would represent a huge effort (15 000 lemmas, +3000 definitions, +3000 notations, etc.)

Goals

For MathComp users:

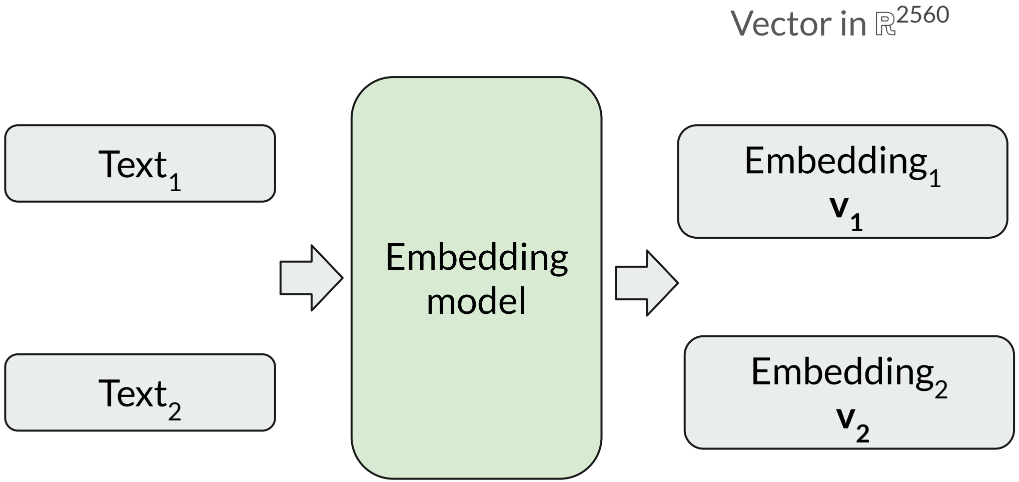
Natural-language search in IDE

For DL community:

Large, expert-reviewed formal <-> informal pairs in Rocq (formalizer/annotator models)

An MCP server to plug LLMs to MathComp (ongoing work: **Crrrocq** with G.Baudart and M. Lelarge)

NL search



If text_1 and text_2 are semantically close then:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 \approx 1$$

If text_1 and text_2 are semantically different then:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 \approx 0$$

VSCode demo

Automatic Evaluation

How to measure the performance of this approach?

Dataset of pairs query/target lemma.

Automatic Evaluation

Extract diverse pairs of theorems/proofs (BM25s)

Stop randomly at some point in proofs, and extract one used statement/definition **not present** in the current file

Given the proof state, and the targeted statement/definition ask an LLM to generate a NL query

```
Lemma nil_class_pgroup (gT : finGroupType) (p : nat) (P : {group gT}) :\n  p.-group P -> nil_class P <= maxn 1 (logn p #|P|).-1.
```

```
move=> pP; move def_c: (nil_class P) => c.
elim: c => // c IHc in gT P def_c pP *; set e := logn p _.
...
by rewrite nil_class_quotient_center ?def_c.
```

```
by rewrite nil_class_quotient_center ?def_c.
```

Query: relationship between nilpotence class of a group and of its quotient by the center

Automatic Evaluation

Query: divisibility of dimensions of vector subspaces

Target lemma: Lemma skew_field_dimS A B : (A <= B)%VS -> \\dim A %| \\dim B.

Target docstring: A lemma stating that if a subalgebra A is contained in a subalgebra B, the dimension of A divides the dimension of B.

Rank: 5

Automatic Evaluation

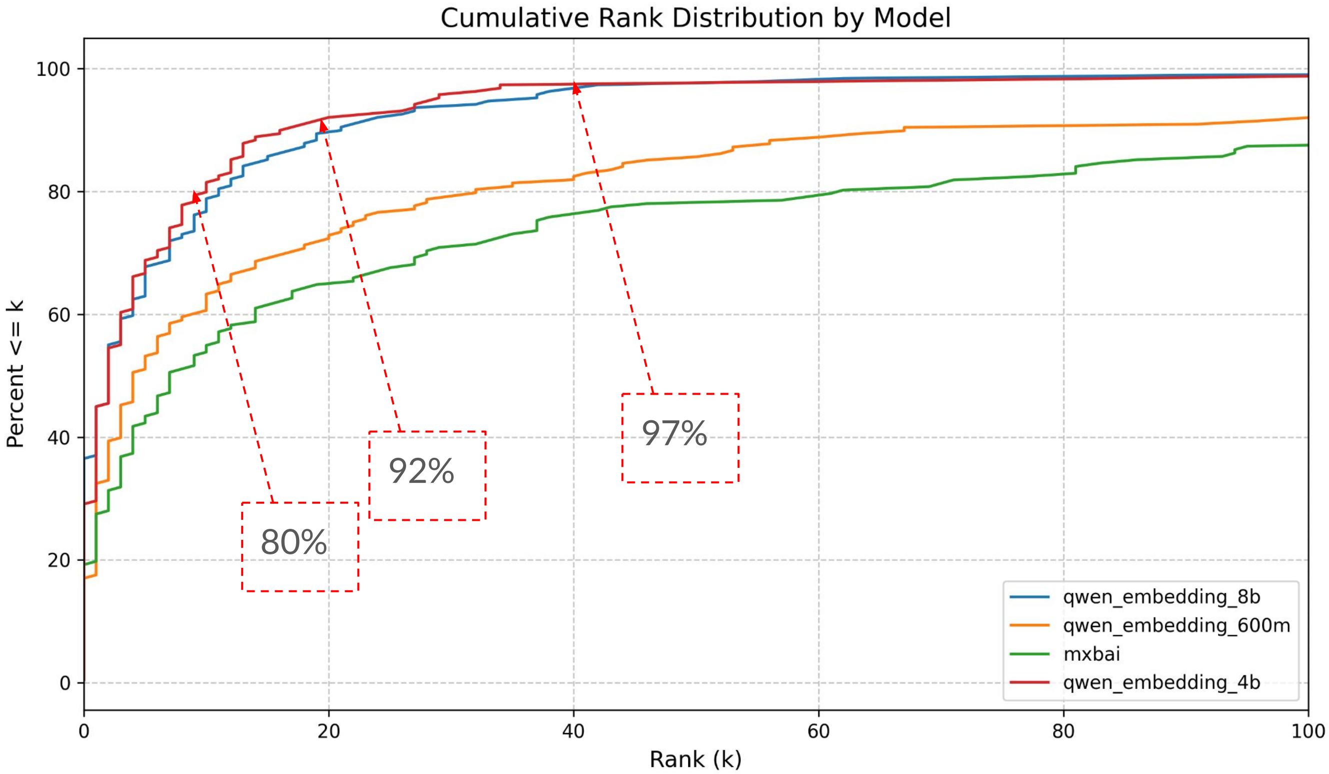
Query: injective function preserves properties of order relations

Targe lemma: Lemma inj_homo : injective f ->\n {homo f : x y / aR x y >-> rR x y} ->\n {homo f : x y / aR' x y >-> rR' x y}.

Target docstring: A lemma stating that an injective function that preserves a relation also preserves the strict version of the relation across the entire domain.

Rank: 9

Automatic Evaluation

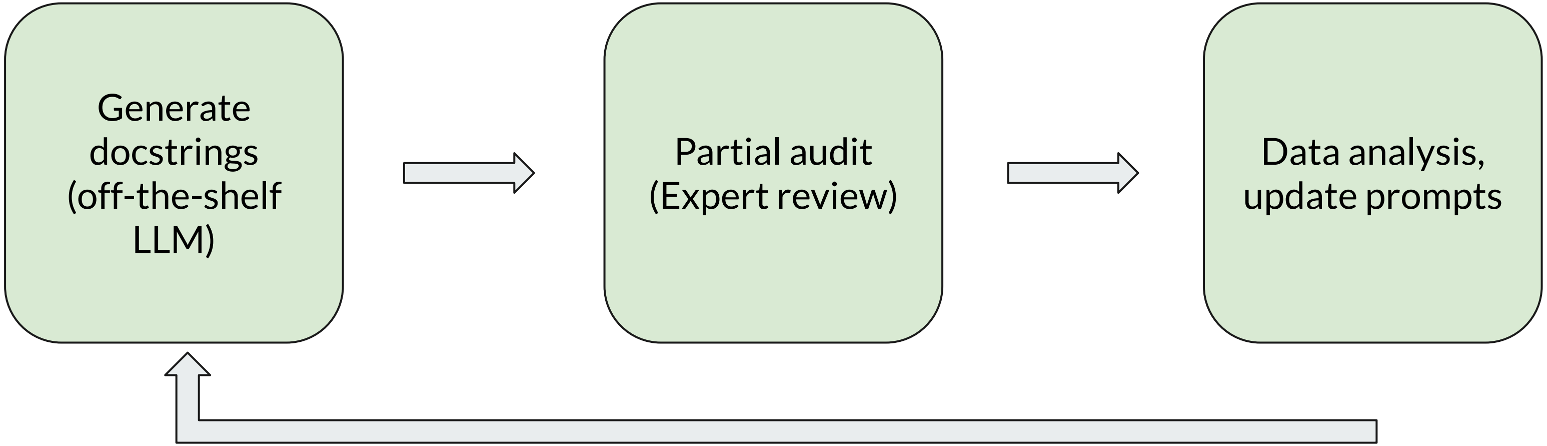


Limitation

We both evaluate the ability of the LLM to formulate “good” queries, the quality of docstrings, and the embedding model.

In practice, multiple queries are probably more efficient than scrolling down dozens of elements.

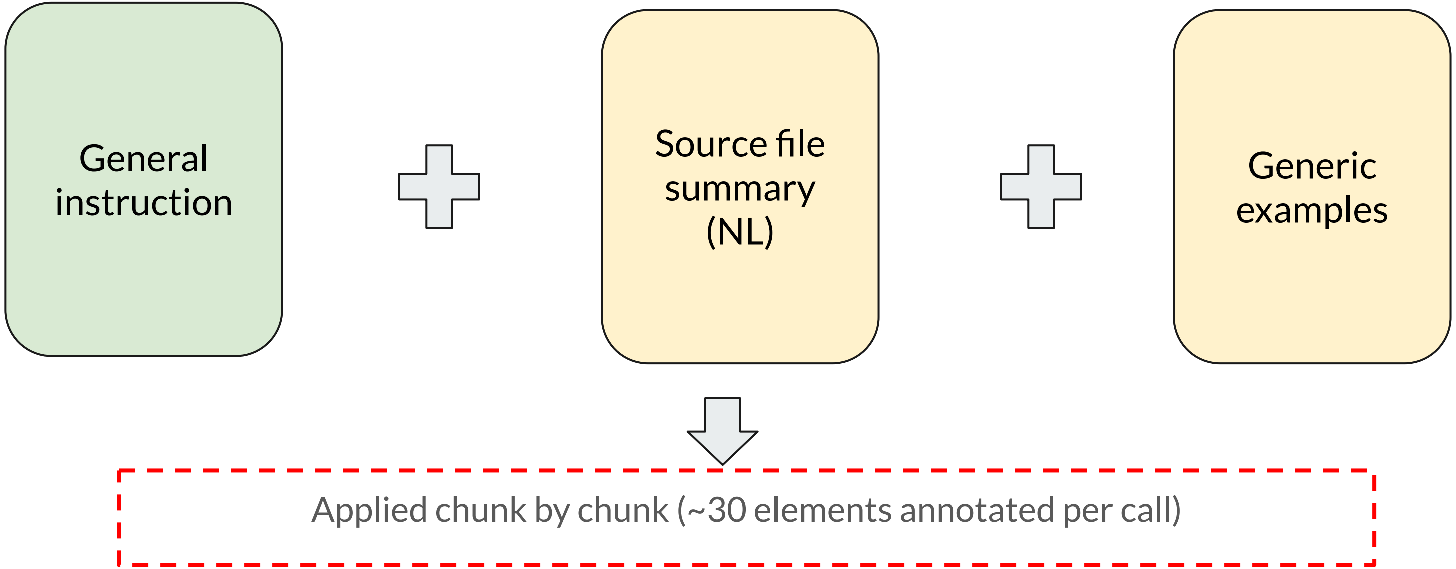
Creating the dataset (ongoing)



Generate docstrings (off-the-shelf LLM)
Partial audit (Expert review)
Data analysis, update prompts

~25€/full run

Step 1: Generate docstrings: version 0



General instruction

Source file summary (NL)

Generic examples

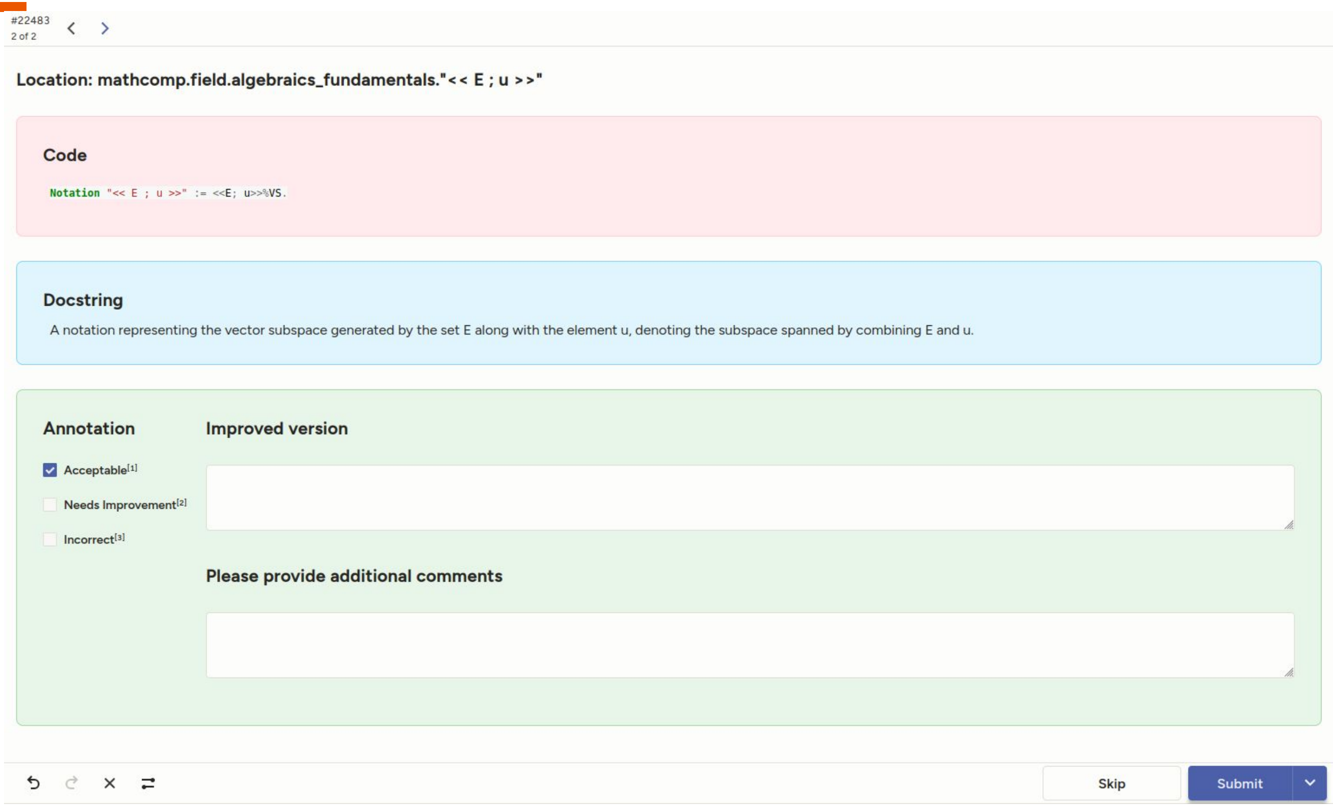
Applied chunk by chunk (~30 elements annotated per call)

Step 2: Expert reviews

Online interface to review docstrings, 3 cases:

- Acceptable
- Needs Improvement
- Incorrect

Acceptable



Location: mathcomp.field.algebraics_fundamentals."<< E ; u >>"

Code

```
Notation "<< E ; u >>" := <<E; u>>%VS.
```

Docstring

A notation representing the vector subspace generated by the set E along with the element u, denoting the subspace spanned by combining E and u.

Annotation

☑ Acceptable[1]

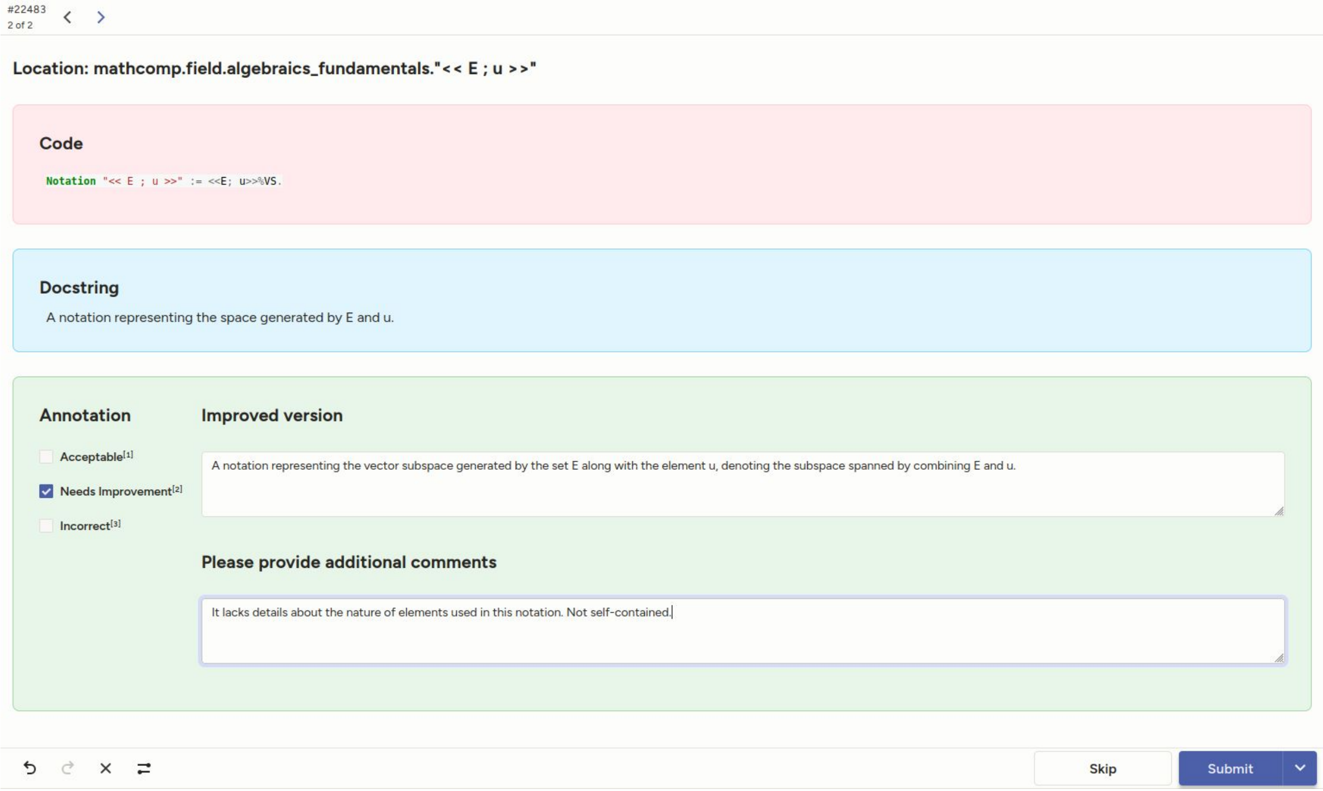
□ Needs Improvement[2]

□ Incorrect[3]

Improved version

Please provide additional comments

Needs improvement



Location: mathcomp.field.algebraics_fundamentals."<< E ; u >>"

Code

```
Notation "<< E ; u >>" := <<E; u>>%VS.
```

Docstring

A notation representing the space generated by E and u.

Annotation

□ Acceptable[1]

☑ Needs Improvement[2]

□ Incorrect[3]

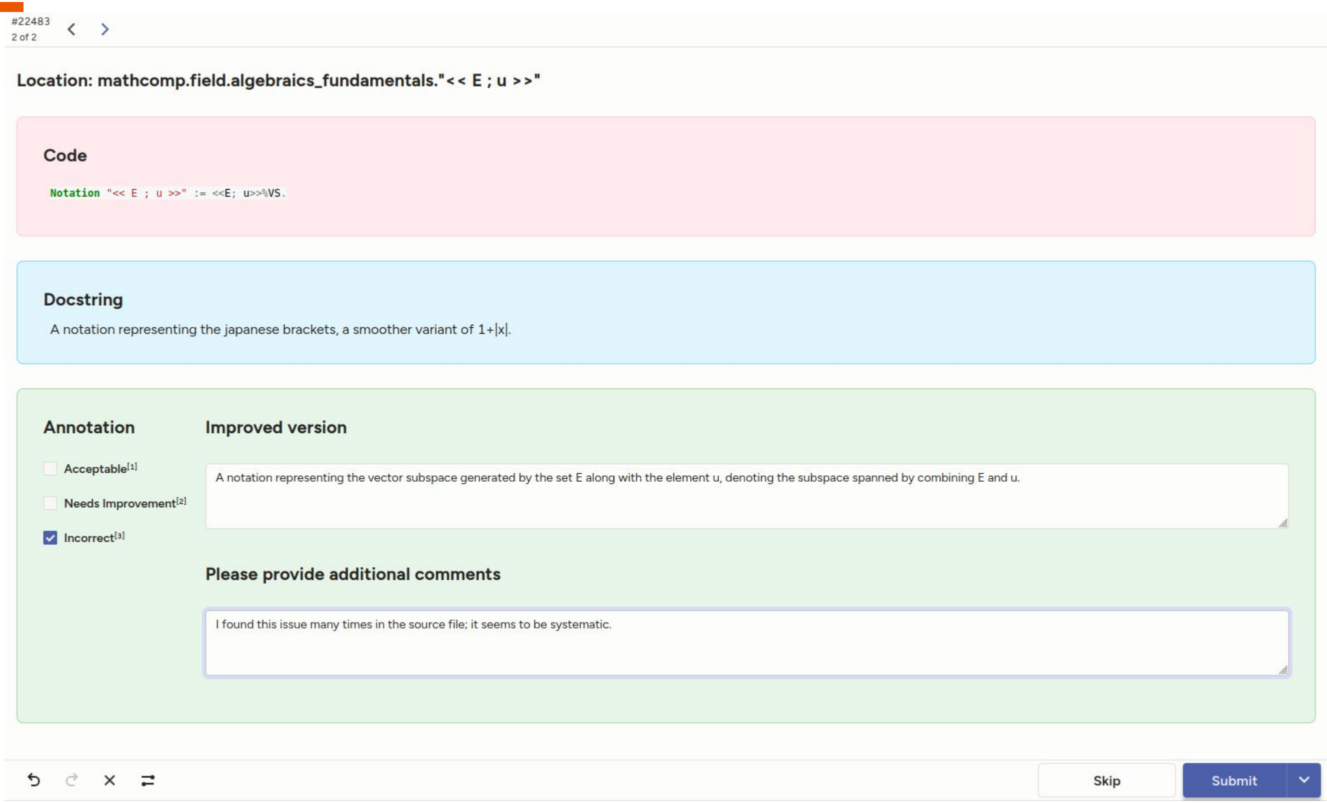
Improved version

A notation representing the vector subspace generated by the set E along with the element u, denoting the subspace spanned by combining E and u.

Please provide additional comments

It lacks details about the nature of elements used in this notation. Not self-contained.

Incorrect



#22483
2 of 2

Location: mathcomp.field.algebraics_fundamentals."<< E ; u >>"

Code

```
Notation "<< E ; u >>" := <<E; u>>%VS.
```

Docstring

A notation representing the japanese brackets, a smoother variant of 1+|x|.

Annotation

□ Acceptable[1]

□ Needs Improvement[2]

☑ Incorrect[3]

Improved version

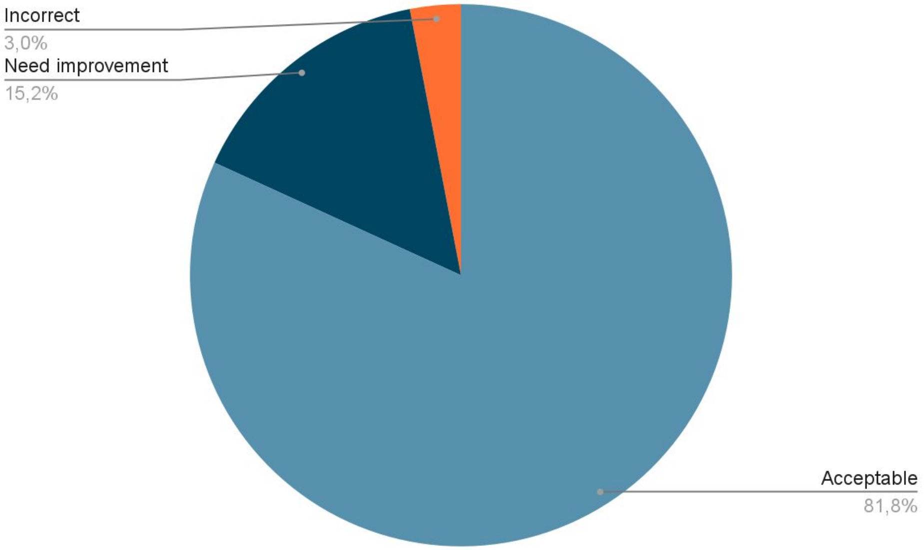
A notation representing the vector subspace generated by the set E along with the element u, denoting the subspace spanned by combining E and u.

Please provide additional comments

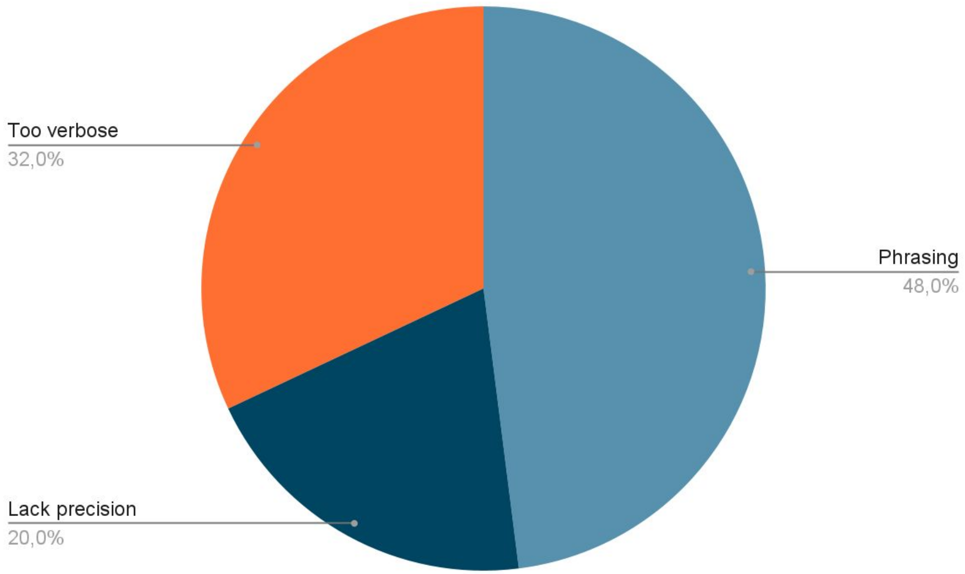
I found this issue many times in the source file; it seems to be systematic.

v0 review (global)

after 600 reviewed docstrings



v0 review (needs improvement)



v1 (ongoing preparation)

- Update each prompt with expert feedback
- New set of rules to have more homogeneity in docstrings form
- ...

Until we reach >95% acceptable docstrings

What’s next

We would obtain a high quality dataset of pairs of formal statement and informal statement

- Train a model to predict docstring (annotator) given file context and formal statement

- Train a model to predict formal statement given file context and docstring.

Thank you!

To contribute to LLM4Docq reach out on rocq-prover.zulipchat.com (@Théo Stoskopf)

Or by mail at: theo.stoskopf@inria.fr

Look at https://github.com/LLM4Rocq/LLM4Docq