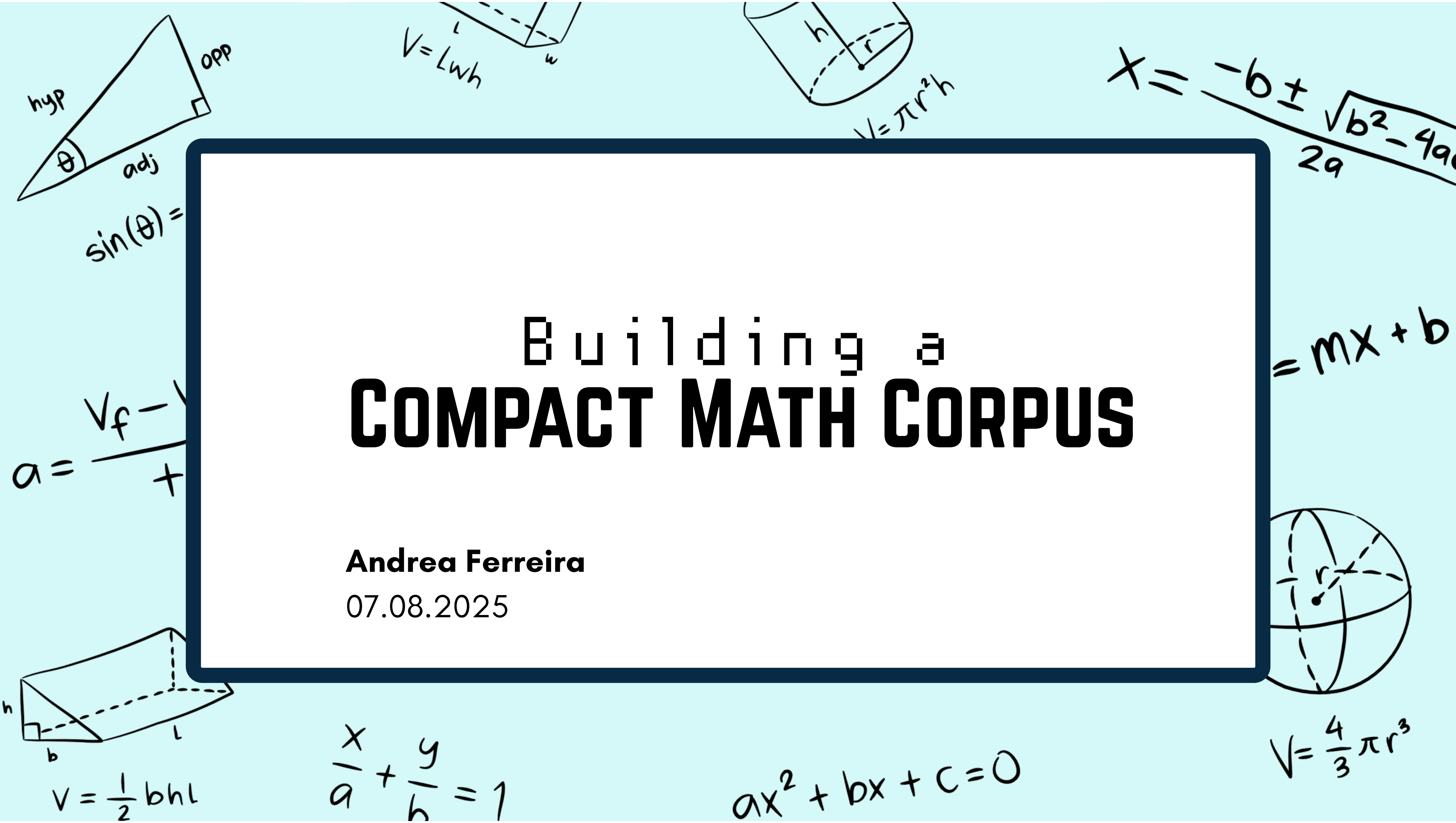


Building a
Compact Math Corpus

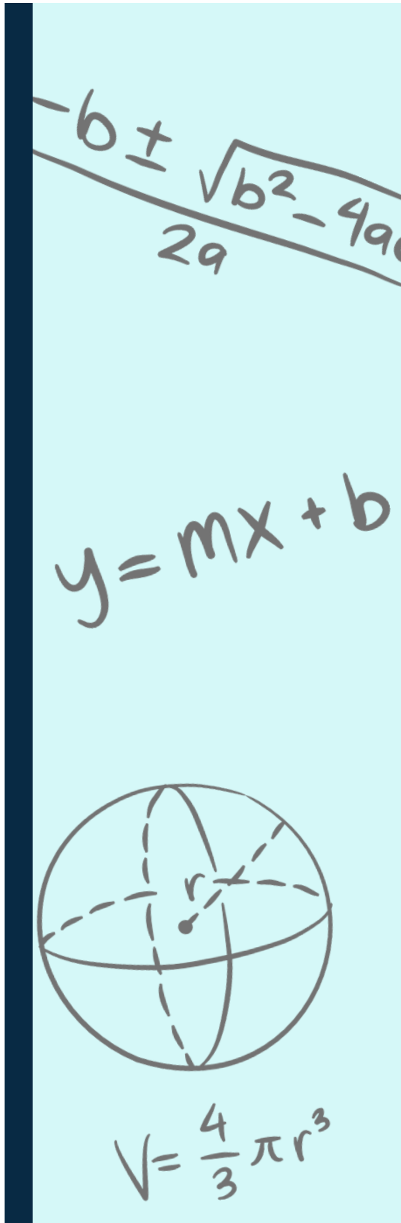
Andrea Ferreira
07.08.2025

Why This Matters?

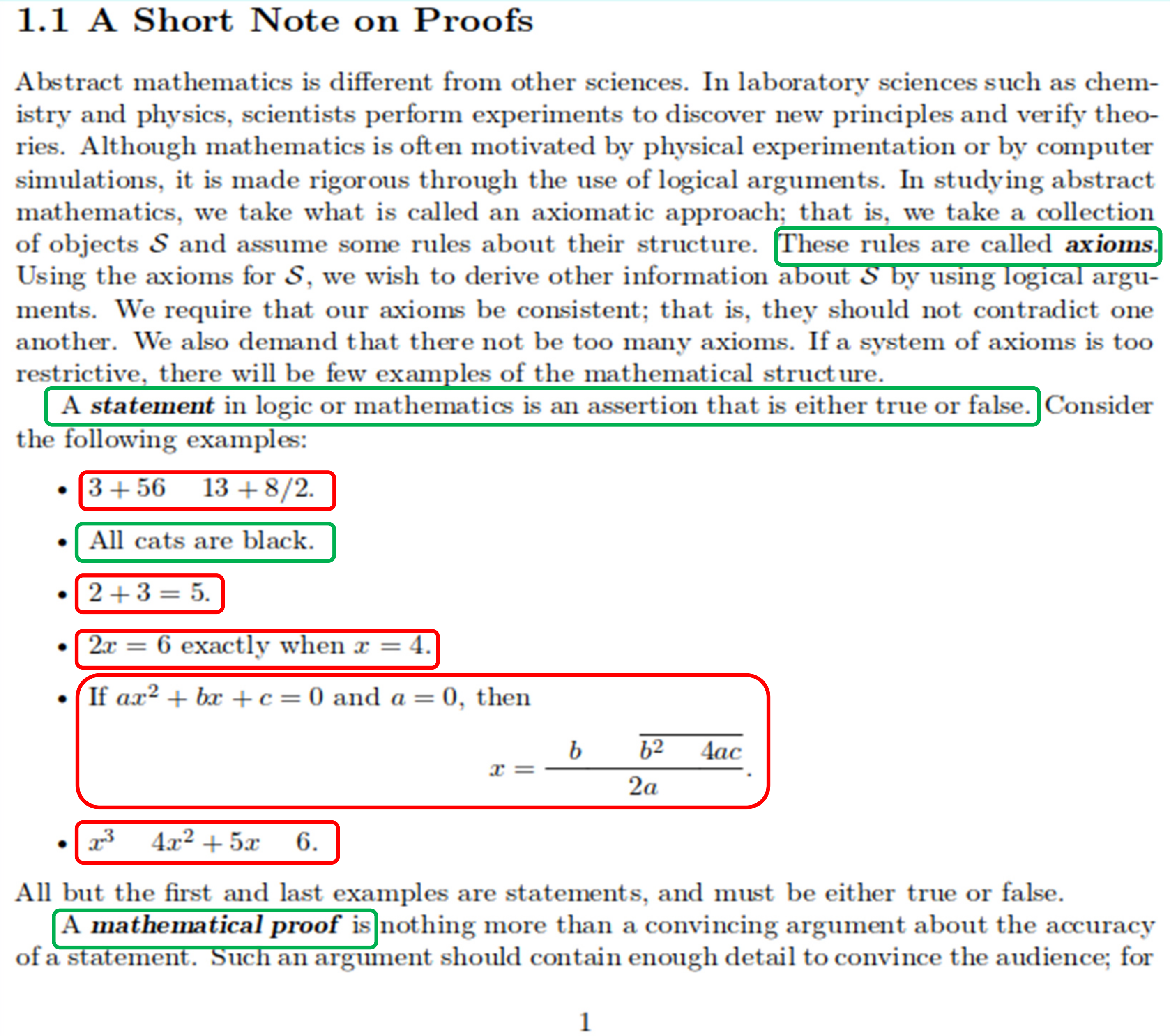
- **Cobbe et al. (2021):** *Mathematical reasoning reveals a critical weakness in modern LMs*
- **Hendrycks et al. (2021):** *Accuracy remains low even with large transformers*
- **Collard et al. (2022):** *Symbols and domain vocabulary challenge general NLP*

The Gap

- Few annotated corpora for math syntax/semantics
- Math NLP is high-impact and underexplored
- We need tools that are replicable, lightweight, and domain-aware



Judson (2022). *Abstract Algebra: Theory and Applications*.



1.1 A Short Note on Proofs

Abstract mathematics is different from other sciences. In laboratory sciences such as chemistry and physics, scientists perform experiments to discover new principles and verify theories. Although mathematics is often motivated by physical experimentation or by computer simulations, it is made rigorous through the use of logical arguments. In studying abstract mathematics, we take what is called an axiomatic approach; that is, we take a collection of objects $\mathcal{S}$ and assume some rules about their structure. These rules are called ***axioms***. Using the axioms for $\mathcal{S}$, we wish to derive other information about $\mathcal{S}$ by using logical arguments. We require that our axioms be consistent; that is, they should not contradict one another. We also demand that there not be too many axioms. If a system of axioms is too restrictive, there will be few examples of the mathematical structure.

A ***statement*** in logic or mathematics is an assertion that is either true or false. Consider the following examples:

- $3 + 56 \quad 13 + 8/2$.
- All cats are black.
- $2 + 3 = 5$.
- $2x = 6$ exactly when $x = 4$.
- If $ax^2 + bx + c = 0$ and $a = 0$, then
$$x = \frac{b \quad \overline{b^2 \quad 4ac}}{2a}.$$
- $x^3 \quad 4x^2 + 5x \quad 6$.

All but the first and last examples are statements, and must be either true or false.

A ***mathematical proof*** is nothing more than a convincing argument about the accuracy of a statement. Such an argument should contain enough detail to convince the audience; for

Objectives

Build Better NLP for Math!

What We Set Out to Do

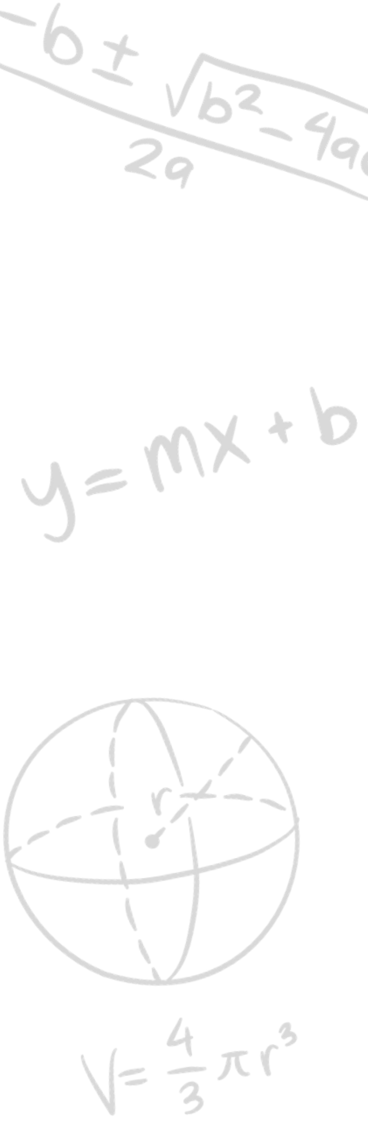
- Create a compact, annotated math corpus (CMC)
- Turn textbook math into structured, searchable data

How We Do It

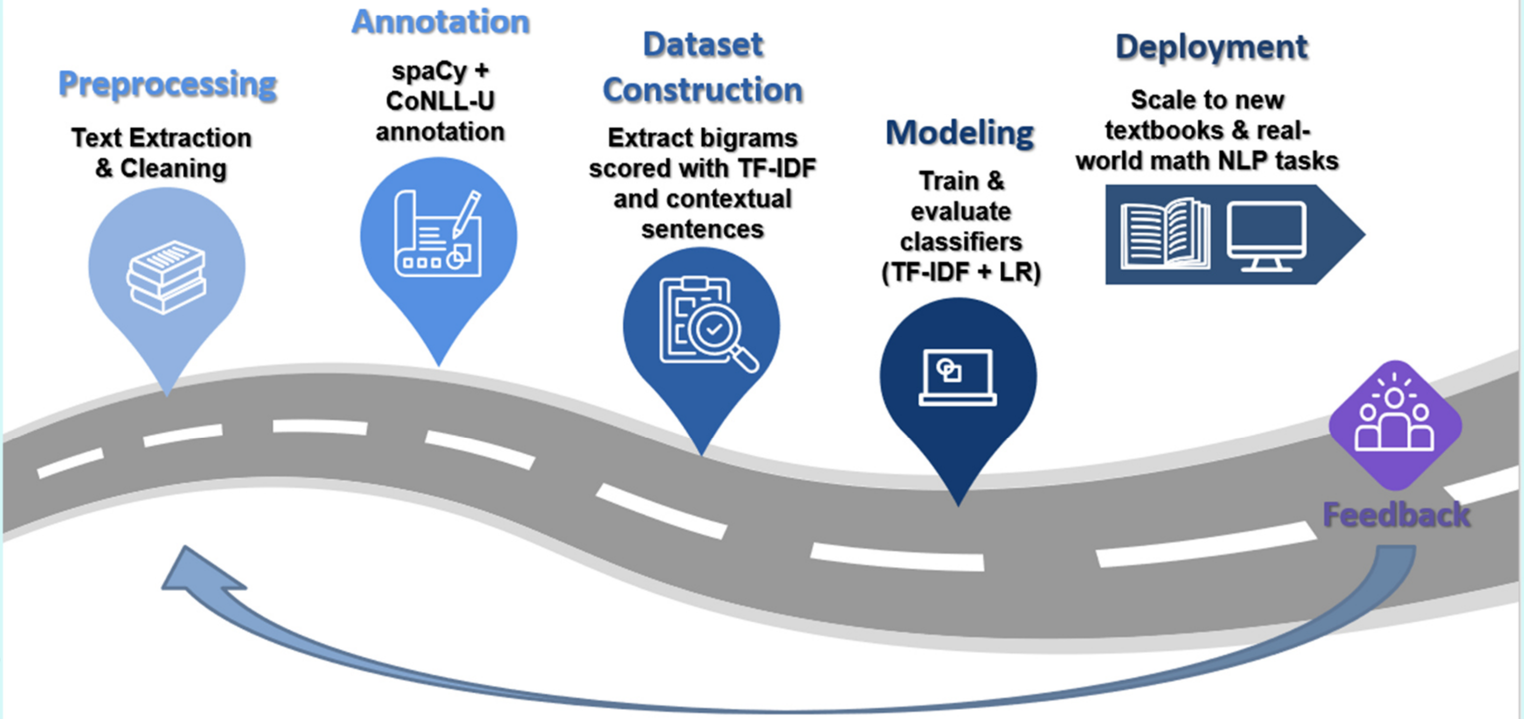
- Use spaCy Small + CoNLL-U for lightweight syntactic annotation
- Extract compounds & MWEs using parsing + TF-IDF, with context sentences

What It Enables

- Concept classification and terminology extraction
- A baseline for low-resource, reproducible math NLP



Build Better NLP for Math!



Preprocessing

Text Extraction & Cleaning

Annotation

spaCy + CoNLL-U annotation

Dataset Construction

Extract bigrams scored with TF-IDF and contextual sentences

Modeling

Train & evaluate classifiers (TF-IDF + LR)

Deployment

Scale to new textbooks & real-world math NLP tasks

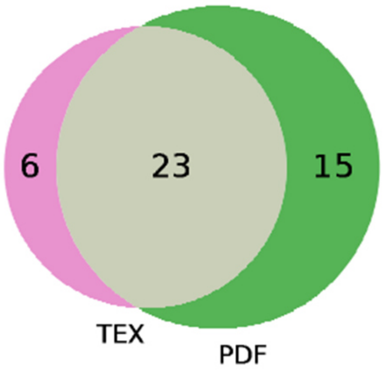
Feedback

Preprocessing

What is Inside the Corpus

- Abstract Algebra (Judson, 2022)
- Linear Algebra (Hefferon, 2022)
- Discrete Math (Levin, 2024)

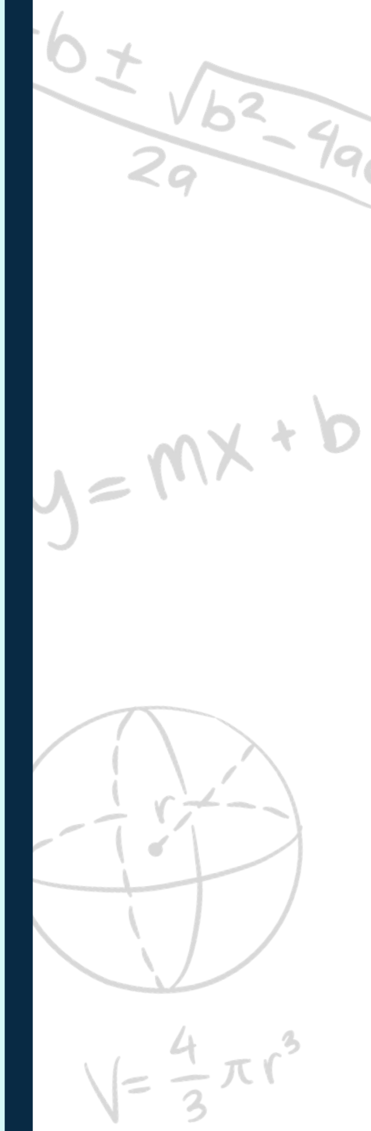
From the
AIM Open Textbook Initiative
https://textbooks.aimath.org/



Trade-offs: PDF vs LaTeX

- **PDF is scalable** and widely used in education
- **LaTeX** offers richer structure

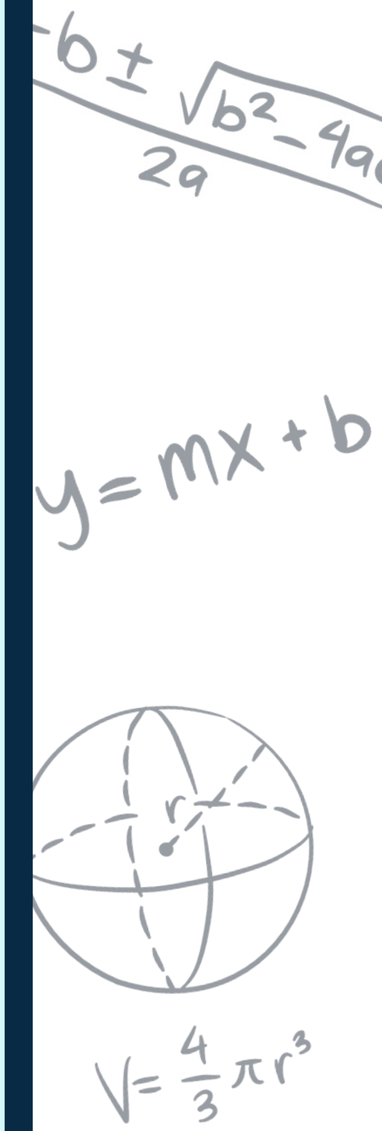
Comparative test showed **PDF** = **viable** for compound extraction



Preprocessing

Corpus Processing Pipeline

- Extract text from **PDFs using PyMuPDF**
- Normalize layout and structure to **JSON**
- **Apply preprocessing:**
    - ➤ Anomaly detection and cleaning steps
    - ➤ Filter non-ASCII text, remove noise
- Annotate using spaCy Small → CoNLL-U format

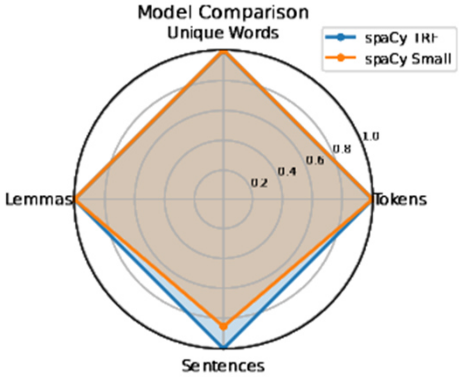


Annotation

Why CoNLL-U Matters

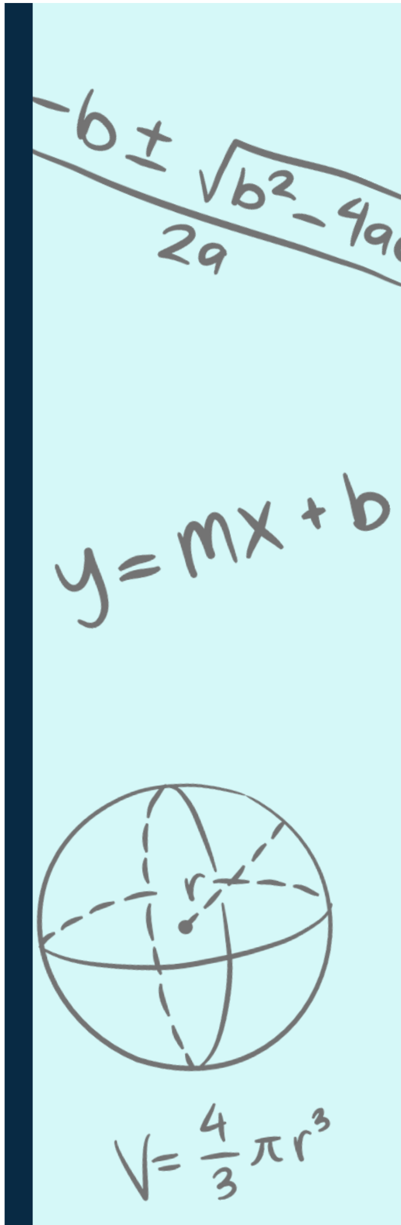
- Detects compounds & MWEs
- Helps de-noise unstructured output
- Boosts TF-IDF ranking of math terms

Bigram	TF-IDF (**Before CoNLL-U**)	TF-IDF (**After CoNLL-U**)
vector space	1328.53	1957.41
= ?1	188.91	noise removed



Trade-offs: spaCy Small vs Transformer

- spaCy **Small** = near-identical token/lemma counts
- spaCy **Transformer** = finer sentence segmentation, no extra payoff in our math NLP pipeline



Dataset

(1) Source Corpora

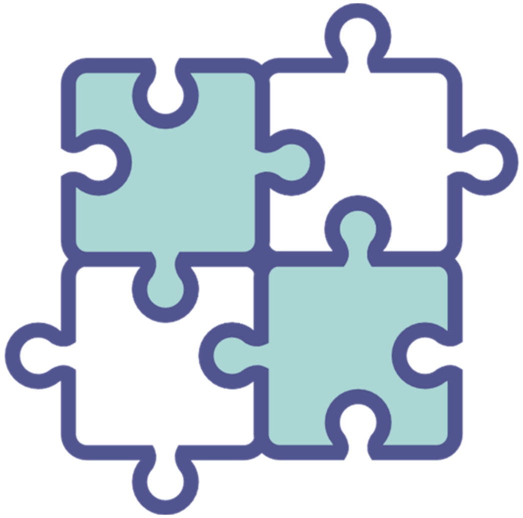
CMC - Math textbook corpus
&
UD English Web Treebank (UD-EWT)

Both in CoNLL-U format

(2) Bigrams Extraction

Extract bigrams

Remove overlaps to isolate domain-specific terms



Bigram: **vector space**
Sentence: "**Every vector space has a basis.**"
Label: **Math ✓**

(3) Mapping and Labeling

Retrieve sentences containing each bigram

Label as:
✓ **Math (CMC)**
✕ **Non-Math (UD-EWT)**

(4) Final Dataset

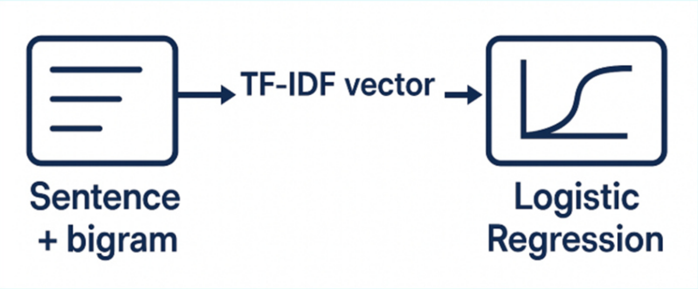
2,796 labeled sentences
Balanced: Math vs. General English

Modeling

Goal

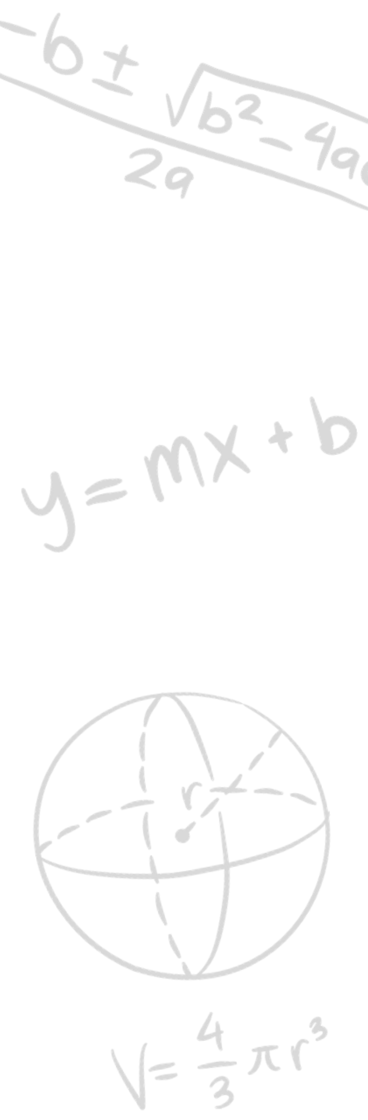
Predict whether a sentence expresses a potential math concept

Smart Classification Task
No AI Overkill



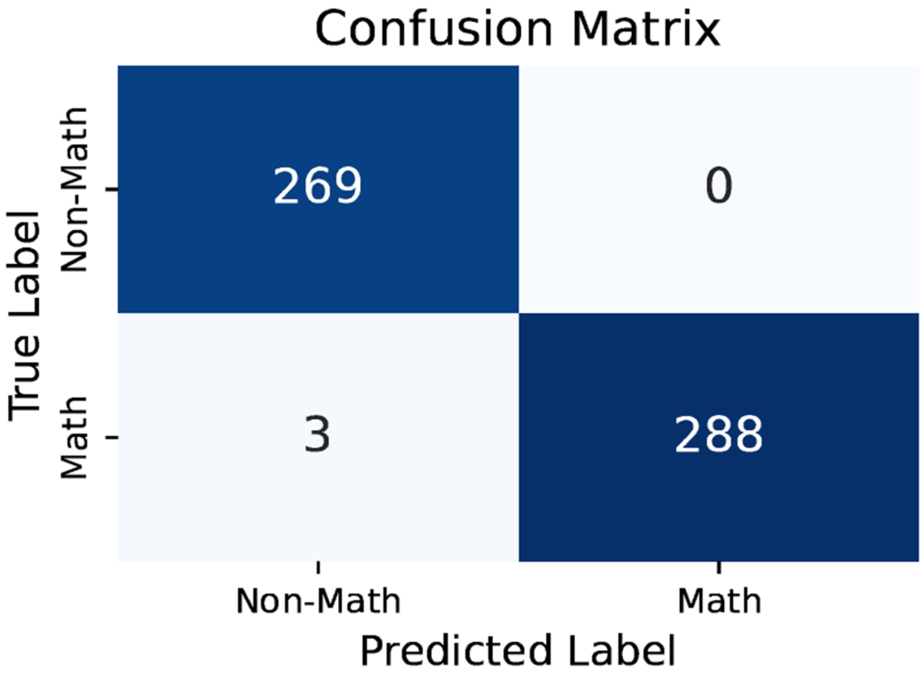
Model Setup

- Binary classification task
- Features: TF-IDF over unigrams + bigrams (max 5,000)
- Classifier: Logistic Regression
- scikit-learn for model implementation



Performance

- Macro F1-score: 0.996 ± 0.003
- 3 false negatives, 0 false positives
- Generalizes to unseen concepts like '*probability distribution*'



Limitations

- Relies on surface-level features (TF-IDF)
- Can struggle with educational terms (e.g., *school project, homework folder*)
- Relies on corpus artifacts (e.g., email IDs)

Zero Shot

Bigram	Sentence	Predicted Class	Probability
probability distribution	*The shape of a probability distribution affects how likely specific outcomes are.*	1 (Math)	81.6%
homework folder	*He forgot his homework folder on the bus.*	0 (Non-Math)	49.3%

Discussion

Why It Works

- Uses syntactic patterns, not deep learning
- Zero-shot generalization to unseen phrases
- Minimal setup = scalable for any topic

Where It Fails

- General-purpose tools struggle with math language
- Parsing errors from PDF source data
- No gold-standard annotations for math
- Compact models trade accuracy for efficiency

Conclusion

- CMC is compact, interpretable, and efficient
- Enables low-resource NLP research in math
- Supports reproducibility and educational use

What is Next?

- **Expand the CMC** with more textbooks
- Create a **gold-standard** set with human-labeled concept terms
- Test generalization across **other domains**
- Integrate symbolic elements for richer, **multimodal modeling**

References

- **Cobbe et al. (2021).** *Training verifiers to solve math word problems.* arXiv:2110.14168.
- **Collard et al. (2022).** *Extracting mathematical concepts from text.* arXiv:2208.13830.
- **Lu et al. (2024).** *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.* International Conference on Learning Representations (ICLR).
- **Hefferon, J. (2022).** *Linear Algebra.* https://hefferon.net/linearalgebra/.
- **Hendrycks et al. (2021).** *Measuring mathematical problem solving with the MATH dataset*. arXiv:2103.03874.
- **Honnibal et al. (2020).** *spaCy: Industrial-strength NLP in Python*. https://spacy.io.
- **Judson (2022).** *Abstract Algebra: Theory and Applications*. http://abstract.pugetsound.edu.
- **Levin, O. (2024).** *Discrete Mathematics.* https://discrete.openmathbooks.org/.
- **Nivre et al. (2016).** *Universal Dependencies v1: A multilingual treebank collection*. Language Resources and Evaluation Conference (LREC).

Thank you for listening!

Andrea Ferreira
Independent Researcher
andreafer.uni@gmail.com

NALOMA - August 2025

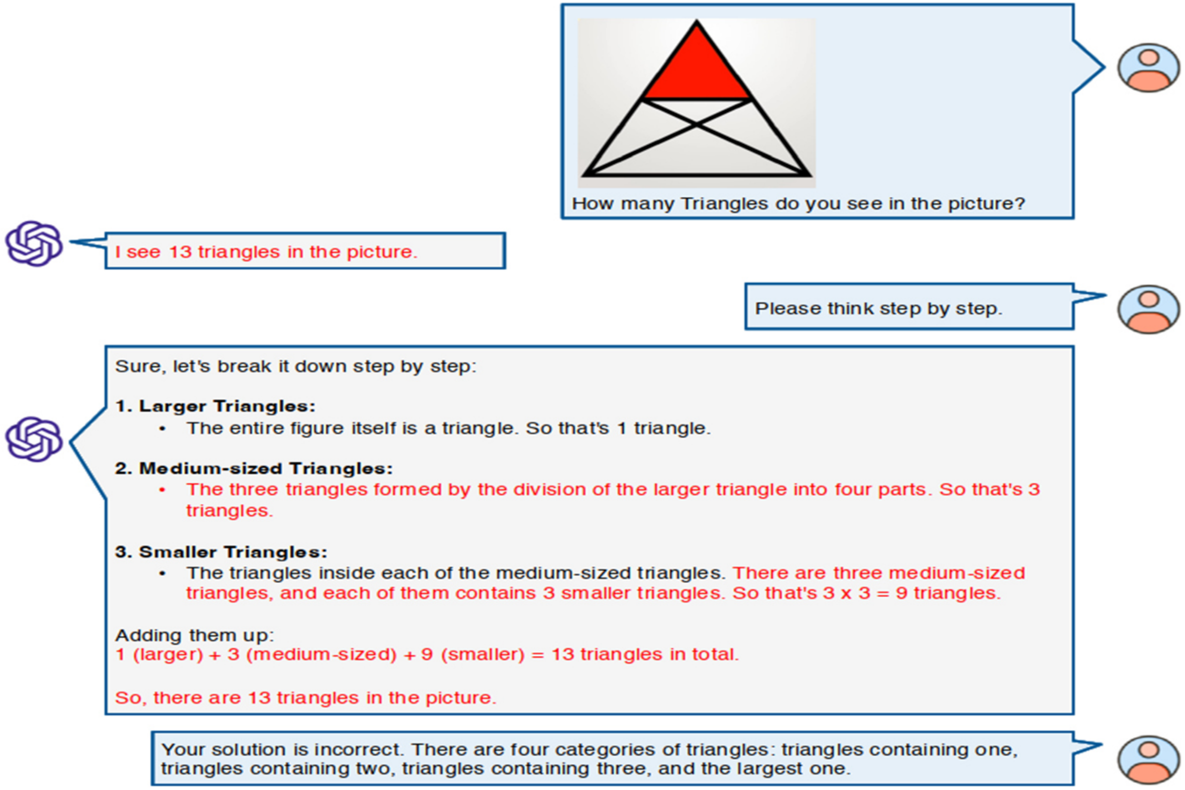
ChatGPT

MATHVISTA

Evaluating Math Reasoning

in Visual Contexts

https://mathvista.githu



How many Triangles do you see in the picture?

I see 13 triangles in the picture.

Please think step by step.

Sure, let's break it down step by step:

1. Larger Triangles:
- The entire figure itself is a triangle. So that's 1 triangle.

2. Medium-sized Triangles:
- The three triangles formed by the division of the larger triangle into four parts. So that's 3 triangles.

3. Smaller Triangles:
- The triangles inside each of the medium-sized triangles. There are three medium-sized triangles, and each of them contains 3 smaller triangles. So that's 3 x 3 = 9 triangles.

Adding them up:
1 (larger) + 3 (medium-sized) + 9 (smaller) = 13 triangles in total.

So, there are 13 triangles in the picture.

Your solution is incorrect. There are four categories of triangles: triangles containing one, triangles containing two, triangles containing three, and the largest one.

CoNLL-U

```
doc_id = 200
sent_id = 216
text = These rules are called axioms.
  These   these   DET DT  Number=Plur|PronType=Dem    2   det _   _
  rules   rule    NOUN    NNS Number=Plur 4   nsubjpass   _   _
  are be  AUX VBP Mood=Ind|Tense=Pres|VerbForm=Fin    4   auxpass _   _
  called  call    VERB    VBN Aspect=Perf|Tense=Past|VerbForm=Part    0   ROOT    _   _
  axioms  axiom   NOUN    NNS Number=Plur 4   oprd    _   SpaceAfter=No
  .   .   PUNCT   .   PunctType=Peri  4   punct   _   SpaceAfter=No
```

D Model Interpetation

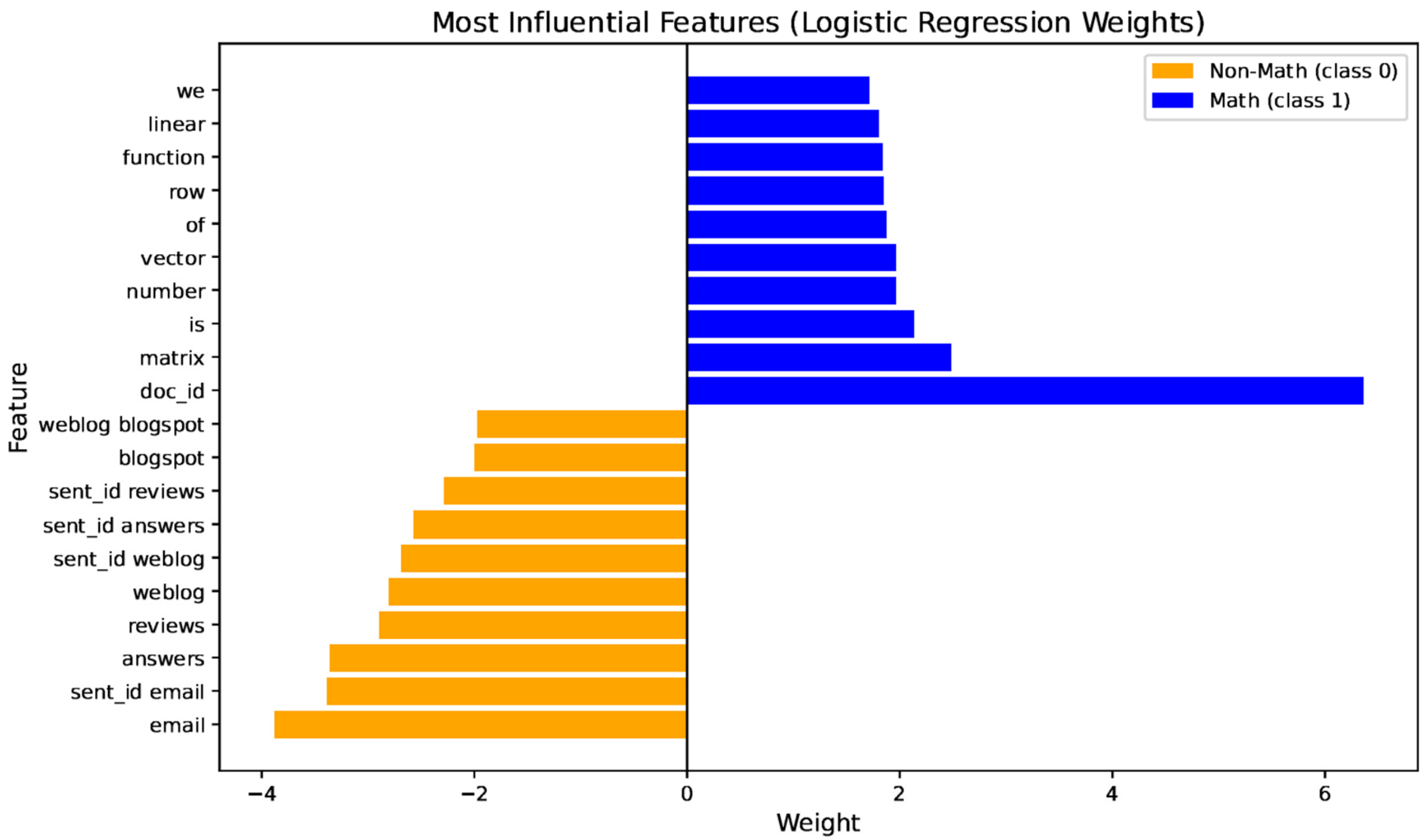


Figure 5: Top weighted features from the logistic regression model. Positive weights (blue) indicate strong association with mathematical concepts, while negative weights (orange) are associated with non-mathematical content. Some features may reflect structural tokens (e.g., *doc_id*, *email*) from the dataset.