DS@GT at CheckThat! 2025: Evaluating Context and Tokenization Strategies for Numerical Fact Verification

Maximilian Heil, Aleksander Pramov

Georgia Tech

Task and Motivation

Objective: Verify claims with numerical quantities and temporal expressions

Motivation: Underrepresentation of numerical claims & specifics of numerical reasoning

Measure: Macro F1-Score of a multiclass classification (True, False, Conflicting)

Languages: English (Spanish, Arabic)

Georgia Tech

Research Questions

RQ 1: Does longer context (3 vs. 9) of retrieved evidence snippets improve the veracity prediction?

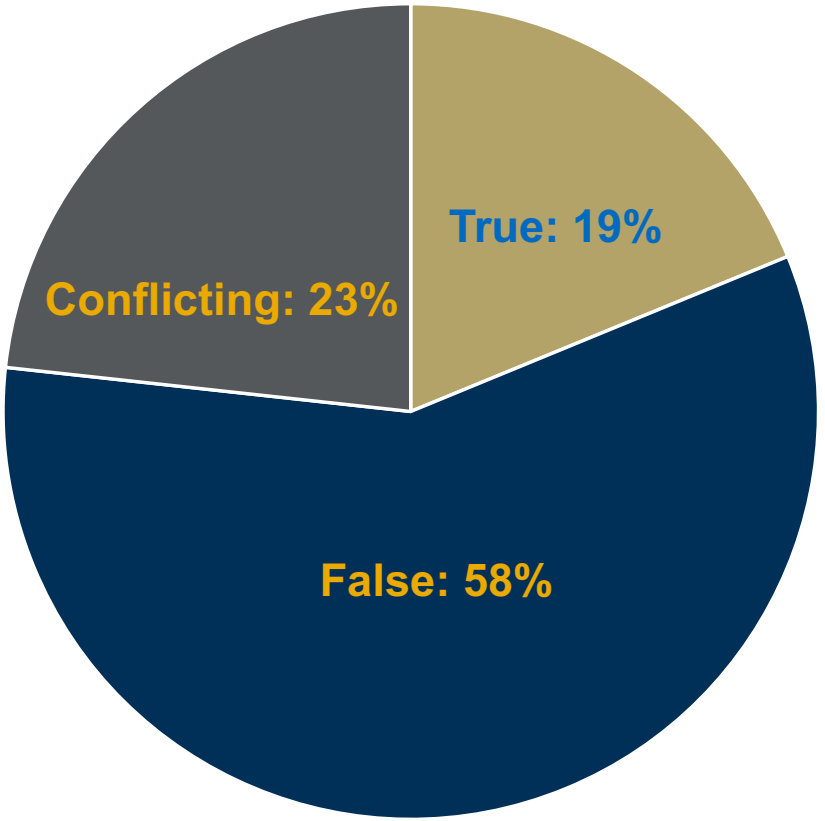
RQ 2: Does R2L-tokenization improve performance?

RQ 3: Does combining long context and R2L-tokenization outperform the other settings?

Data

15,514 Claims

Example: ”The city of Columbus would save $41 million a year if employees had to contribute to their own, guaranteed-check pensions.”



Georgia Tech

Data

15,514 Claims

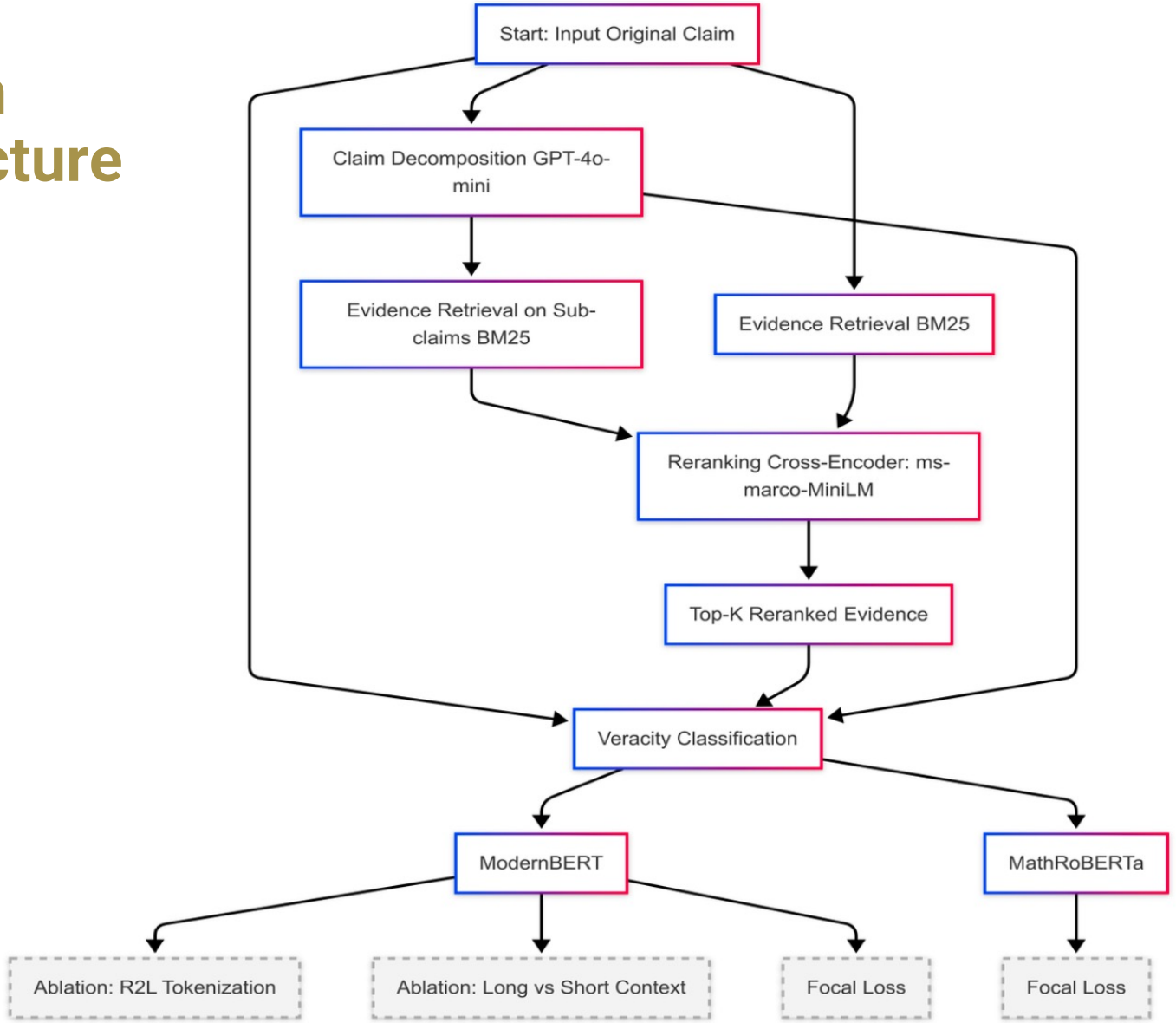
Example: ”The city of Columbus would save $41 million a year if employees had to contribute to their own, guaranteed-check pensions.”

432,320 Evidence Snippets

1. 31 ago 2022 life expectancy at birth for women in the united states dropped 0.8 years from 79.9 years in 2020 to 79.1 in 2021, while life expectancy for ...
2. apr 28, 2014 1.7% of the world's water is frozen and therefore unusable.1. approximately 400 billion gallons of water are used in the united states per day.1.
3. …

Georgia Tech

Solution Architecture



Georgia Tech

Short vs. Long Context for Veracity Classification

Short Context

- 1 Claim
- 3 Questions
- 1 Evidence per question

Context Window: 256

Long Context

- 1 Claim
- 3 Questions
- **3 Evidence per question**

Context Window: 1,024

Georgia Tech

Number Tokenization: Standard L2R vs R2L

Source: https://huggingface.co/spaces/huggingface/number-tokenization-blog

Georgia Tech

Results

Table 1
Experiment and Ablation Study Results

	Run	Train		Validation				
		Macro-a. F1	Acc.	Macro-a. F1	False F1	Conflicting F1	True F1	Acc.
	Benchmark	0.75	0.67	0.56	0.79	0.48	0.41	0.66
	Our-Data	0.56	0.67	0.52	0.80	0.29	0.46	0.64
	Short-Context	0.50	0.70	0.52	0.77	0.42	0.37	0.61
RQ 1	Long-Context	0.64	0.74	0.52	0.78	0.37	0.41	0.62
RQ 2	R2L Short-Context	0.38	0.60	0.45	0.79	0.40	0.16	0.63
RQ 3	R2L Long-Context	0.42	0.60	0.47	0.79	0.32	0.30	0.62
	Submission	**0.63**	**0.71**	**0.57**	**0.81**	**0.36**	**0.55**	**0.66**
	PEFT	0.49	0.63	0.49	0.78	0.30	0.37	0.63
	Focal-Loss	0.65	0.75	0.57	0.81	0.41	0.50	0.67

Georgia Tech

Conclusion & Future Research Avenues

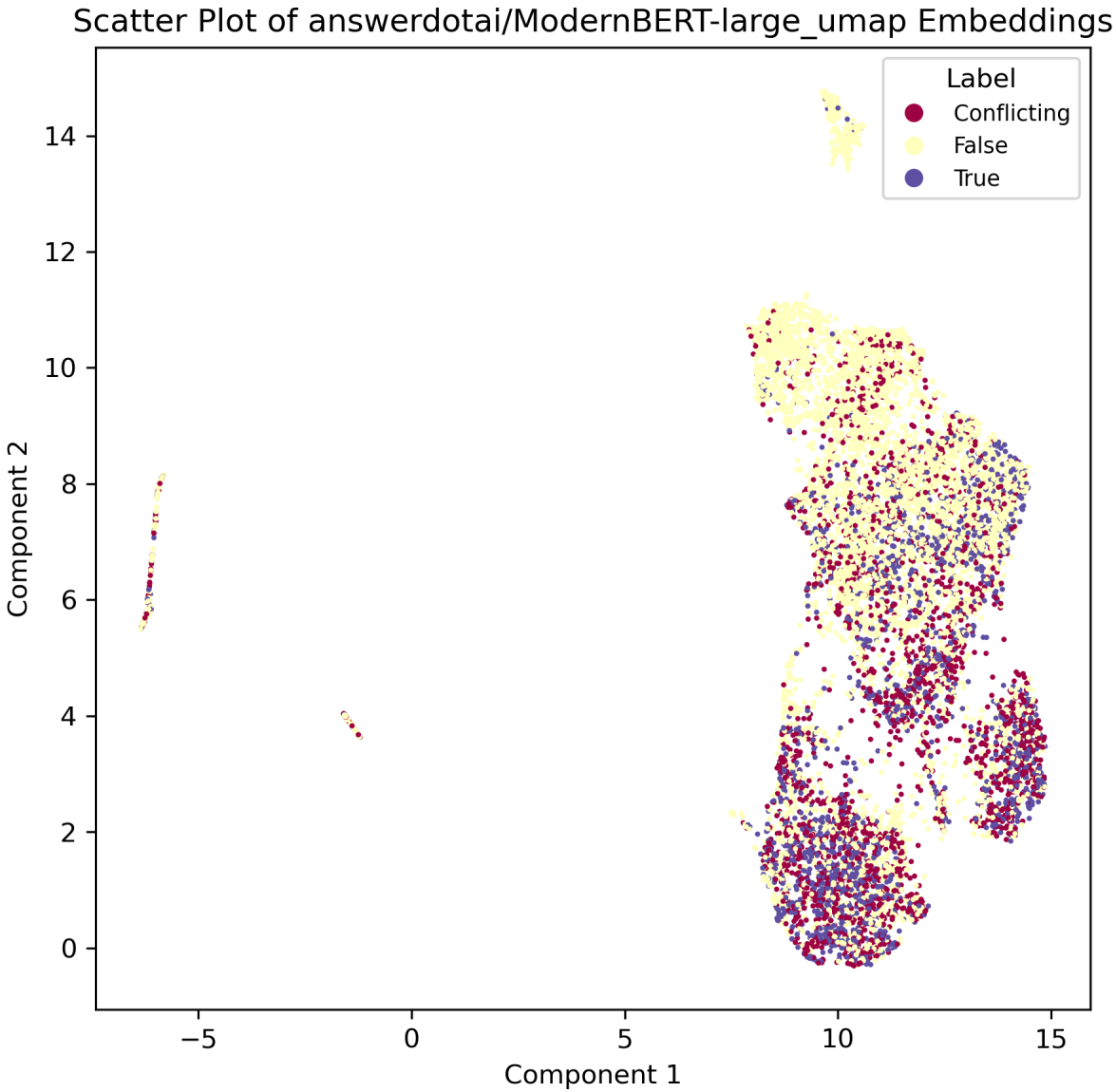
- **Information Retrieval: Dense**
- Number-sensitive model-selection: single digits or unique token models
- Classification: Ensemble
- Normalizing Numbers and Dates

Georgia Tech

Thank you!

Georgia Tech

Data


Scatter Plot of answerdotai/ModernBERT-large_umap Embeddings
Label
Conflicting
False
True
Component 2
Component 1
14
12
10
8
6
4
2
0
−5
0
5
10
15