NAACL 2025

Fine-grained Fallacy Detection with Human Label Variation

Alan **RAMPONI**,[1] Agnese **DAFFARA**,[2,3] Sara **TONELLI**[1]

[1]Fondazione Bruno Kessler [2]University of Pavia [3]University of Stuttgart

Fallacies are *Arguments that seem valid but are not* (used *intentionally* or *unintentionally*)

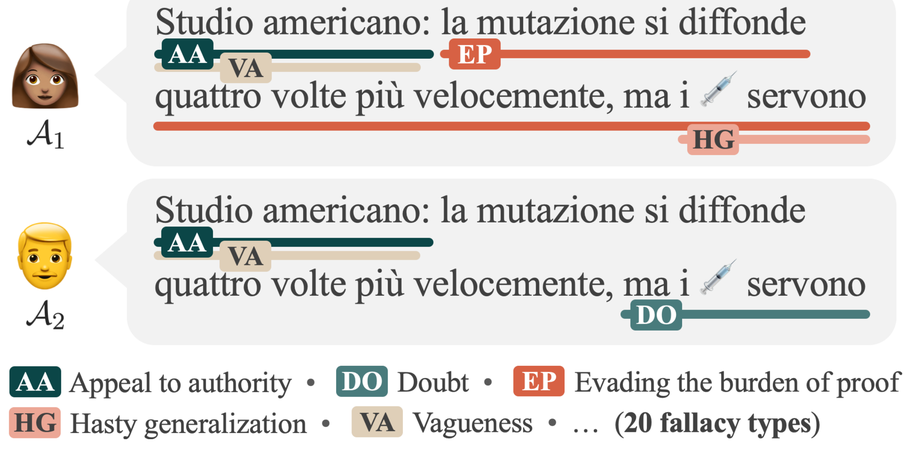
– Aristotle

Recognizing fallacies in everyday argumentation plays a key role in **developing individuals' critical thinking skills**, contributing to **mitigate faulty and harmful argumentation at its root**

DATASET

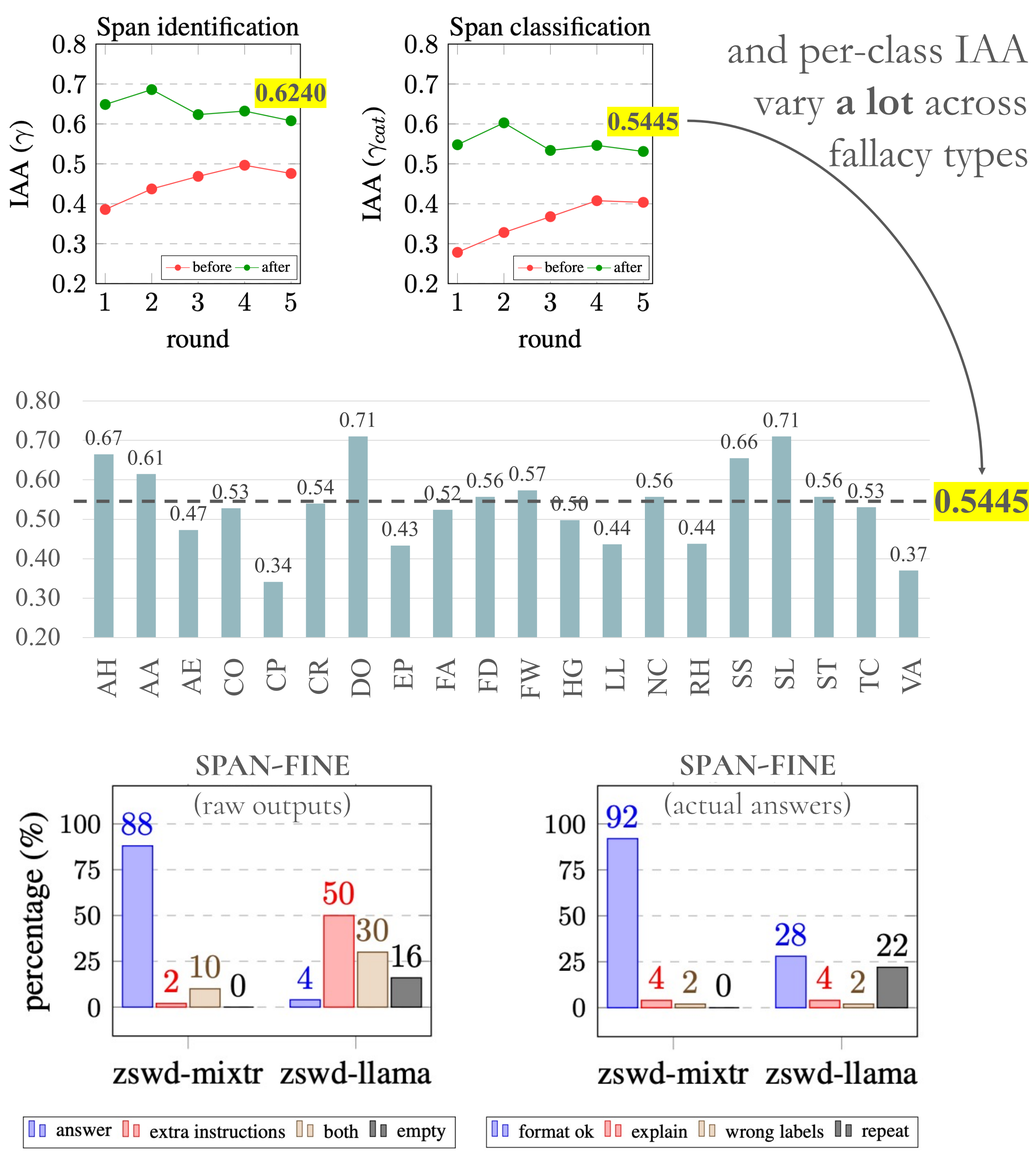
FAINA: The first dataset for fine-grained fallacy detection with human label variation

- Language/genre: 🇮🇹 Italian, 🌐 social media posts
- Multiple gold standards: 🤝 *genuine* disagreement!
- Fine-grained annotation: span-level w/ overlaps
- Large class inventory: 20 fallacy types
- Large time coverage: ⏳ 4 years (2019–2022)
- Multiple topics: public discourse on 🔄 *migration*, 🌱 *climate change*, and 🏥 *public health*



ANALYSIS

Disagreement *is not* noise!



Accounting for human label variation

	macro labels (3)	*all labels* (20)
post level	POST-COARSE	POST-FINE
span level	SPAN-COARSE	SPAN-FINE

micro F1 · span F1 *with overlaps*

⚠ **Individual** test set scores are then macro-averaged

EXPERIMENTS

SETUP

multi-task learning models

MVML model — *|A| multi-label decoders, each returning all the labels exceeding τ*

MVMD model — *|A×F| decoders, each returning the BIO tag for each label and annotation version*

with encoder: **ALB**(erto) or **UMB**(erto)

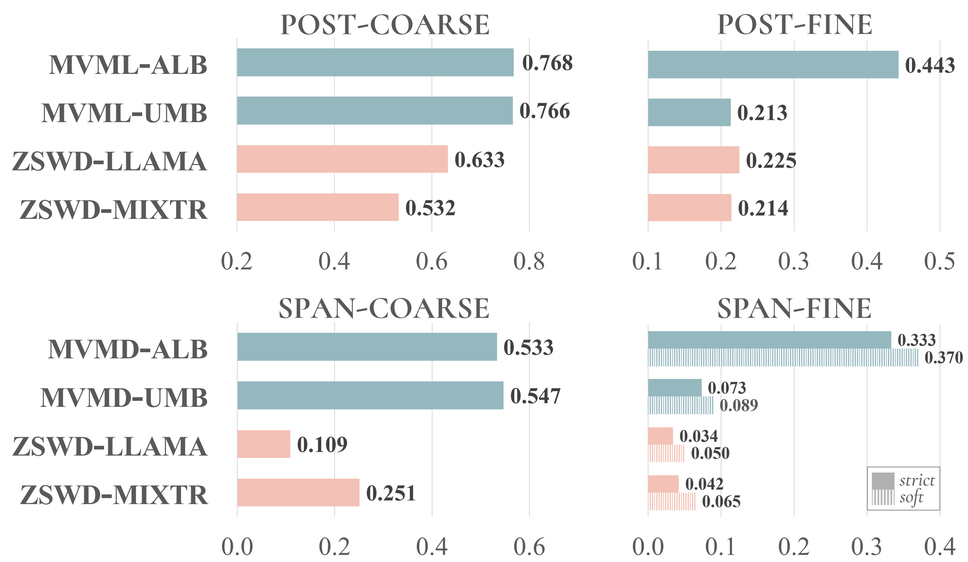
LLMs

LLAMA-3 8B

MIXTRAL 8x7B

ZSWD *zero-shot w/ descriptions*

RESULTS



Interested in this topic? Join the **FADEIT shared task** at EVALITA 2026!