NAACL 2025 – Albuquerque, NM, USA

Fine-grained fallacy detection with human label variation

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

[1] Digital Humanities unit, Fondazione Bruno Kessler, Italy
[2] Department of Humanities, University of Pavia, Italy
[3] IMS, University of Stuttgart, Germany

Fallacies

Arguments that seem valid but are not

– Aristotle

Hasty generalization

Alice got the flu after the influenza vaccine. Vaccines are really useless.

Either used ***intentionally*** (for persuading) or ***unintentionally***

- Not only logical: also *structural*, from *diversion*, due to *language use* [1]
- Difficult to spot: closely follow the patterns of valid arguments [2]
- Impactful: can mislead a wide audience → spread of misinformation

[1] ***Tindale, 2007***. *"Fallacies and Argument Appraisal". Critical Reasoning and Argumentation; CUP.*
[2] ***Musi et al., 2022***. *"Developing Fake News Immunity: Fallacies as Misinformation Triggers During the Pandemic". OJCMT.*

Fallacy detection

Why is it useful?

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

NLP can support this, but current datasets have some limitations:

- Coarse-grained annotations (e.g., *text-level*), few fallacy types, or few annotations
- One fallacy for each text or no overlaps between fallacies in the *span-level* case
- Single ground truth: genuine disagreement is not taken into account!

🐻 FAINA dataset

Dataset for **fine-grained** fallacy detection with **human label variation**

- Focuses on **Italian** social media posts

Studio americano: la mutazione si diffonde quattro volte più velocemente, ma i 💉 servono

en: *American study: mutation spreads four times faster, but 💉 are needed*

🐻 FAINA dataset

Dataset for **fine-grained** fallacy detection with **human label variation**

- Focuses on **Italian** social media posts
- Embraces **multiple plausible answers** and **natural disagreement**

👩🏽 $\mathcal{A}_1$: Studio americano: la mutazione si diffonde quattro volte più velocemente, ma i 💉 servono

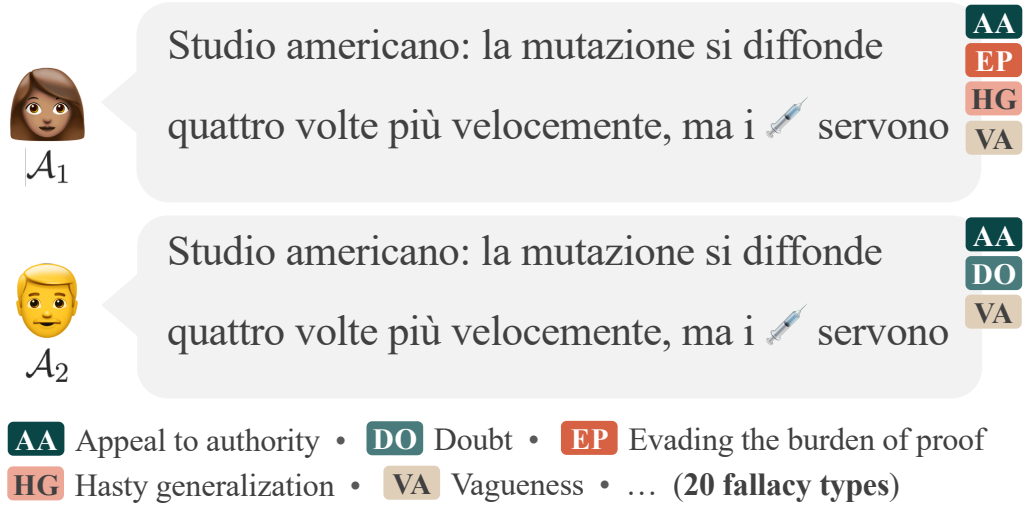
👱 $\mathcal{A}_2$: Studio americano: la mutazione si diffonde quattro volte più velocemente, ma i 💉 servono

en: *American study: mutation spreads four times faster, but 💉 are needed*

🐻 FAINA dataset

Dataset for **fine-grained** fallacy detection with **human label variation**

- Focuses on **Italian** social media posts
- Embraces **multiple plausible answers** and **natural disagreement**
- Large inventory of **20 fallacy types**

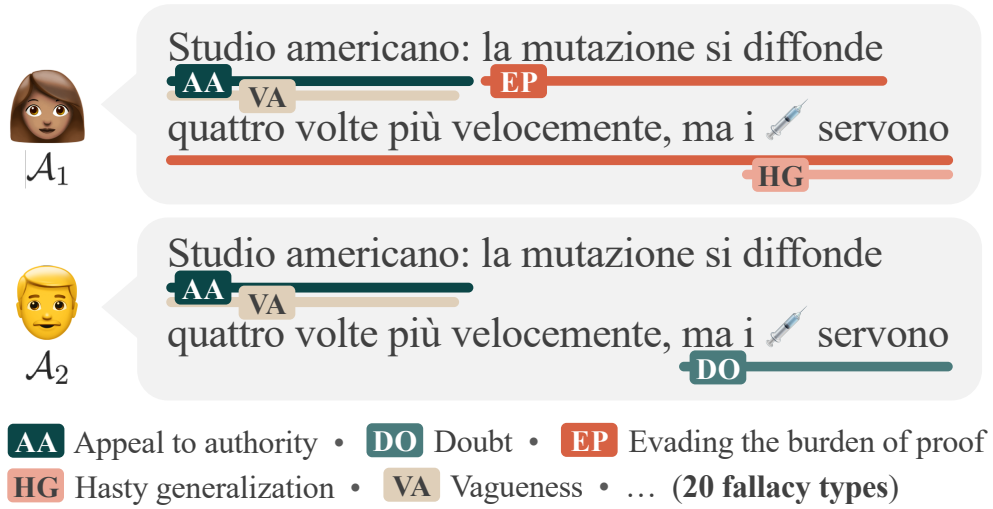


en: *American study: mutation spreads four times faster, but 💉 are needed*

🐻 FAINA dataset

Dataset for **fine-grained** fallacy detection with **human label variation**

- Focuses on **Italian** social media posts
- Embraces **multiple plausible answers** and **natural disagreement**
- Large inventory of **20 fallacy types**
- Fine-grained annotation at the **span-level** with potential **overlaps**



en: *American study: mutation spreads four times faster, but 💉 are needed*

🐻 FAINA dataset: *data collection* and *sampling*

Public discourse on Twitter **minimizing temporal & topic biases**

- **Multi-year**: ⏳ 4-year time frame (2019-01 — 2022-12)
- **Multi-topic**: 🔄 migration, 🌱 climate change, and 🏥 public health
  - *Manually-curated list of 436 neutral keywords derived from trustable glossaries and manuals*

Posts with highest impact to society **minimizing author bias**

- **Top-*k* posts** (*k*=10) by like+retweet [3] for each month/topic
- **Resample posts** by the same authors after their most impactful one

1,440 posts
58,490 tokens

[3] ***Nakov et al., 2022***. "*Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets*". *CLEF*.

🐻 FAINA dataset: *(manual) data annotation*

Annotation goal

Minimize annotation errors whilst keeping signals of human label variation

Intrisically difficult task: fallacy nuances, inventory, granularity, overlaps

⚠ Crowdsourcing is not suitable in this context – discussion is paramount!

- **2 expert annotators** with different sociodemographics and background
- **5 rounds** of annotation/discussion (resolve *errors*, keep *genuine disagreement*)

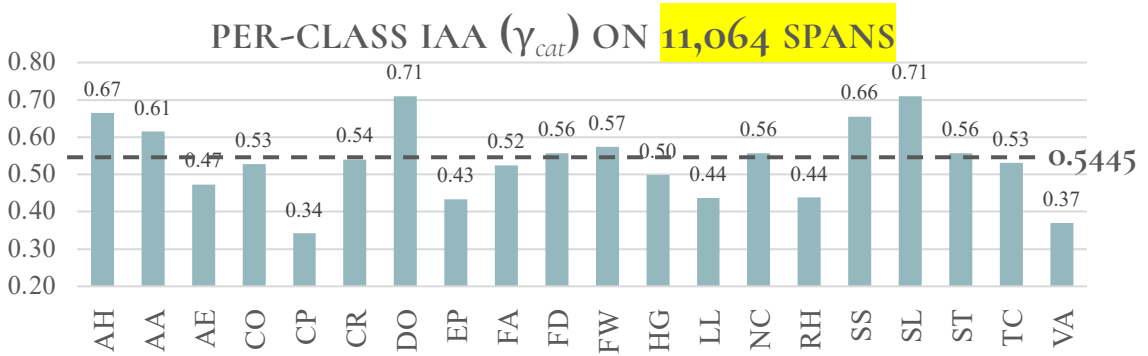
380
person-hours

FAINA dataset: *inter-annotator agreement*

We use γ and γ_{cat} [4, 5] measures for *span-level* labels with *overlaps*

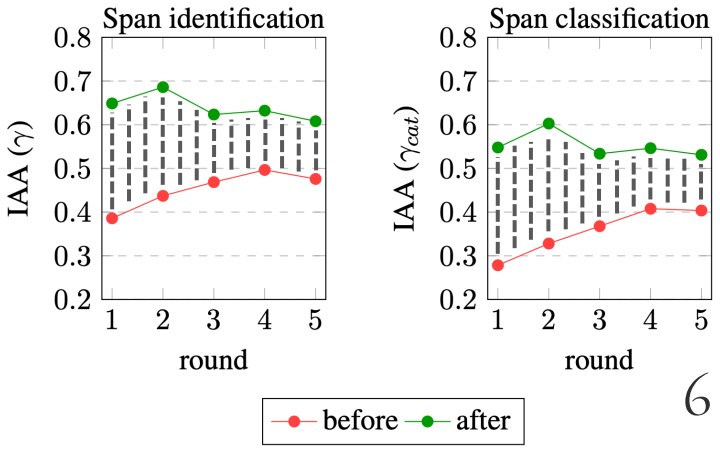
$\gamma = 0.6240$ *(identification)*

$\gamma_{cat} = 0.5445$ *(classification)*



IAA **over rounds** and **before/after discussions**

- Discussions are necessary for such a complex task



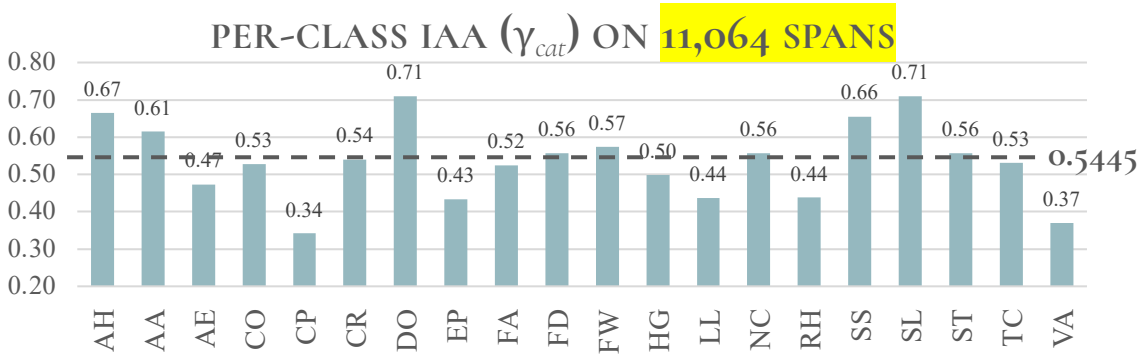
[4] ***Mathet et al., 2015***. "*The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment*". *CL.*
[5] ***Mathet, 2017***. "*The Agreement Measure γ_{cat} a Complement to γ Focused on Categorization of a Continuum*". *CL.*

FAINA dataset: *inter-annotator agreement*

We use γ and γ_{cat} [4, 5] measures for *span-level* labels with *overlaps*

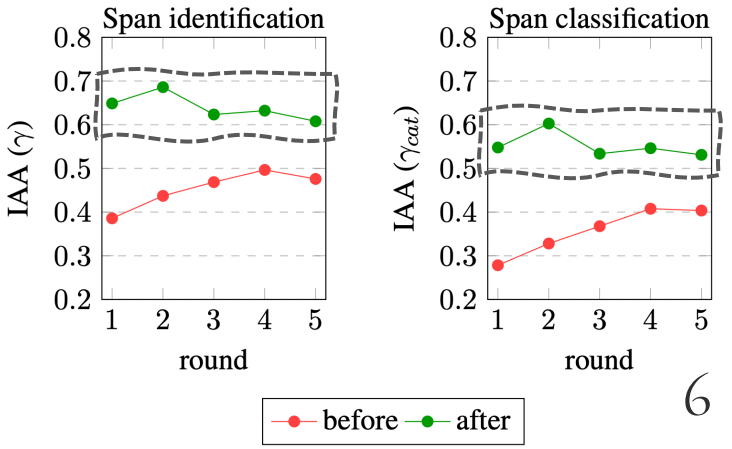
$\gamma = 0.6240$ *(identification)*

$\gamma_{cat} = 0.5445$ *(classification)*



IAA **over rounds** and **before/after discussions**

- Discussions are necessary for such a complex task
- Disagreement can be resolved only partially



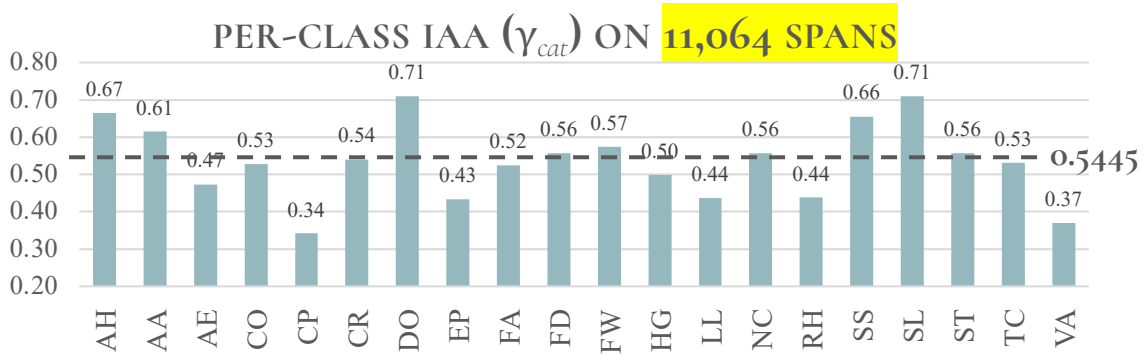
[4] ***Mathet et al., 2015***. "*The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment*". *CL.*
[5] ***Mathet, 2017***. "*The Agreement Measure γ_{cat} a Complement to γ Focused on Categorization of a Continuum*". *CL.*

FAINA dataset: *inter-annotator agreement*

We use γ and γ_{cat} [4, 5] measures for *span-level* labels with *overlaps*

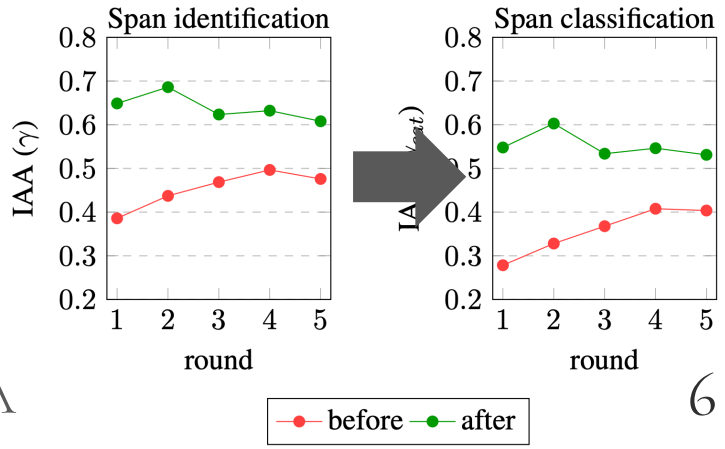
$\gamma = 0.6240$ *(identification)*

$\gamma_{cat} = 0.5445$ *(classification)*



IAA **over rounds** and **before/after discussions**

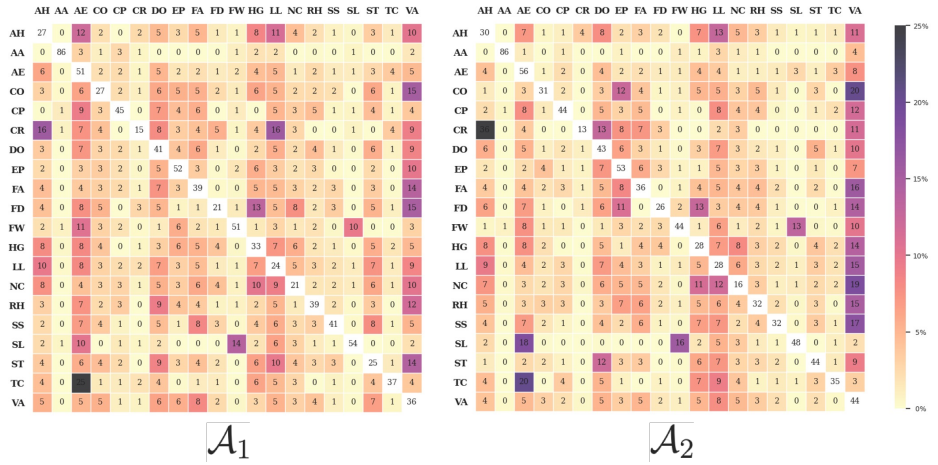
- Discussions are necessary for such a complex task
- Disagreement can be resolved only partially
- Identifying fallacy spans is the main bottleneck for IAA



[4] ***Mathet et al., 2015***. "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment". *CL.*
[5] ***Mathet, 2017***. "The Agreement Measure γ_{cat} a Complement to γ Focused on Categorization of a Continuum". *CL.*

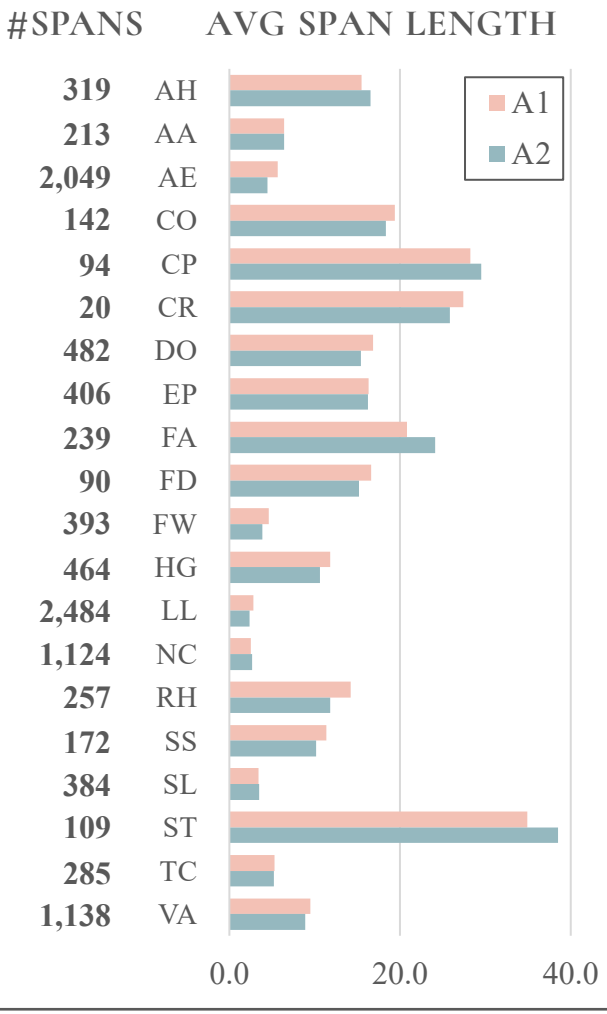
FAINA dataset: *statistics*

- **11,064 spans** – $5{,}532_{\pm 253}$ spans/annotator
- **Avg token length of 7.6**$_{\pm 9.3}$ – from 2.5 [LL] to 36.3 [ST]
- **Dense annotation** – $3.8_{\pm 0.2}$ spans/post
- **Frequent overlaps** – up to $23\%_{\pm 2\%}$ for TC and AE

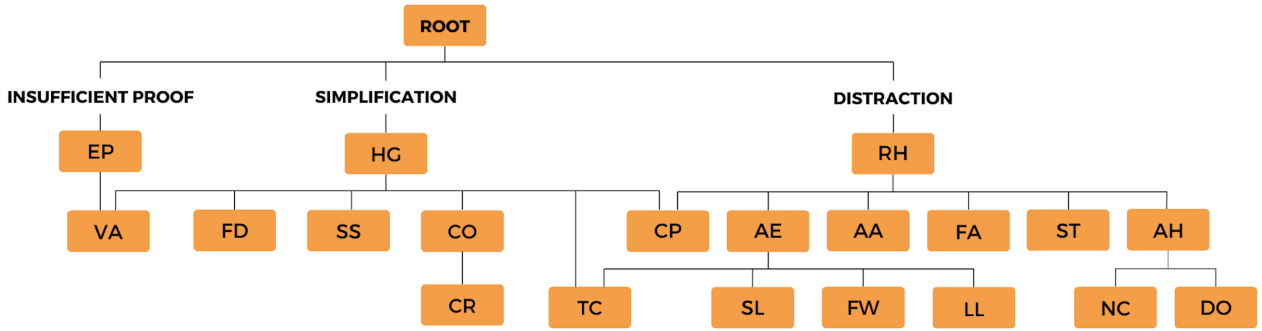


$\mathcal{A}_1$ $\mathcal{A}_2$

Check the paper
for pairwise overlaps!

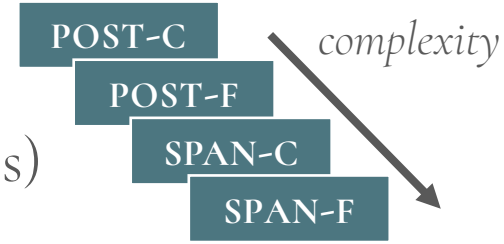


Experiments: *setup*



We cast fallacy detection into **different tasks across two dimensions**:

- **Annotation unit**: *post* level (POST) and *span* level (SPAN)
- **Classification granularity**: *coarse* (C; 3 types) and *fine* (F; all 20 types)



This implies the use of **different evaluation strategies**:

Individual scores are macro-averaged

- **Metrics**: micro F1 for POST, span-level F1 with overlaps [6] for SPAN
- **Modes**: *strict* and *soft* – the latter accounts for varying severity of labeling errors
  - i.e., partial credit (0.5) if the predicted label is an immediate parent of the actual label

[6] ***Da San Martino et al., 2019***. "*Fine-Grained Analysis of Propaganda in News Articles*". *EMNLP*.

Experiments: *models*

Modeling goal

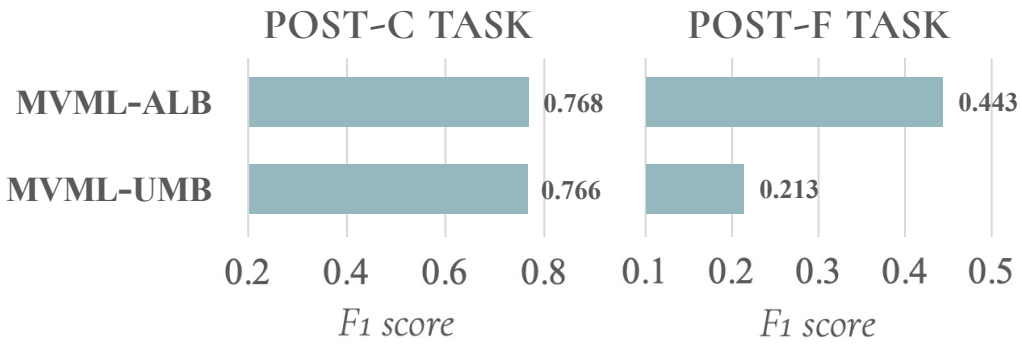
Account for human label variation in fallacy detection

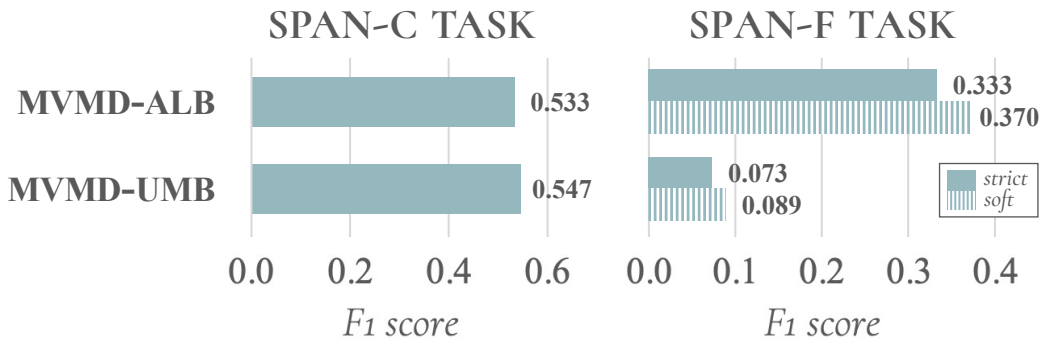
Multi-task learning to jointly model signals of *individual annotations*

- MVML (*multi-view, multi-label*) model for **POST tasks** — MVML-ALB, MVML-UMB
  - |A| multi-label decoders, each outputting all labels exceeding a threshold τ
- MVMD (*multi-view, multi-decoder*) model for **SPAN tasks** — MVML-ALB, MVML-UMB
  - |A×F| decoders, each outputting the BIO tag for each label & annotation version

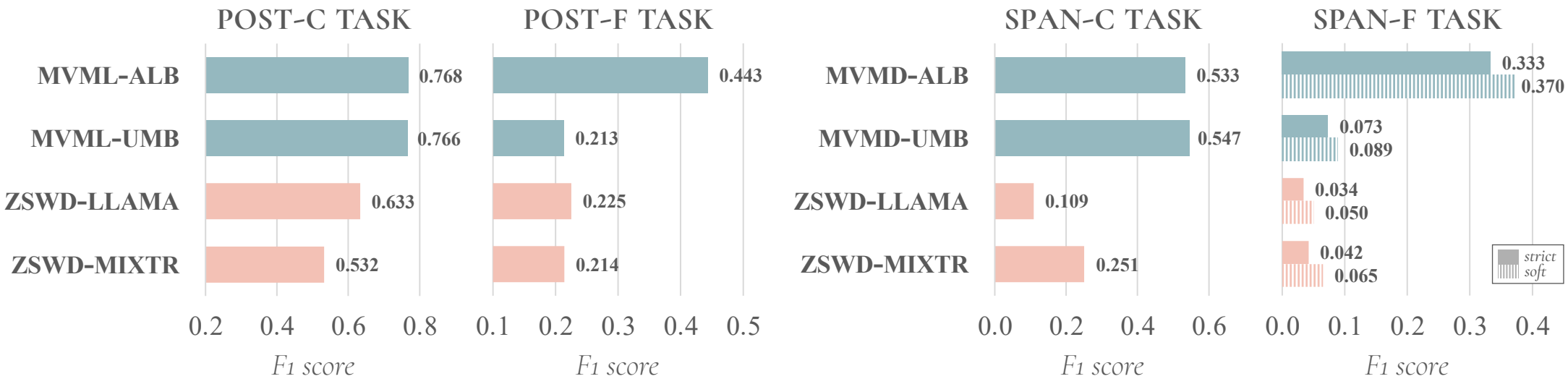
***Shared encoders**: widespread models pretrained on Italian data (ALBERTO & UMBERTO)*

Results


POST-C TASK
POST-F TASK
MVML-ALB
MVML-UMB
0.768
0.766
0.443
0.213
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4
0.5
F1 score
F1 score


SPAN-C TASK
SPAN-F TASK
MVMD-ALB
MVMD-UMB
0.533
0.547
0.333
0.370
0.073
0.089
strict
soft
0.0
0.2
0.4
0.6
0.0
0.1
0.2
0.3
0.4
F1 score
F1 score

Results



? *To what extent we can expect to challenge more traditional approaches with instruction-tuned LLMs in a zero-shot (ZS) setup?*

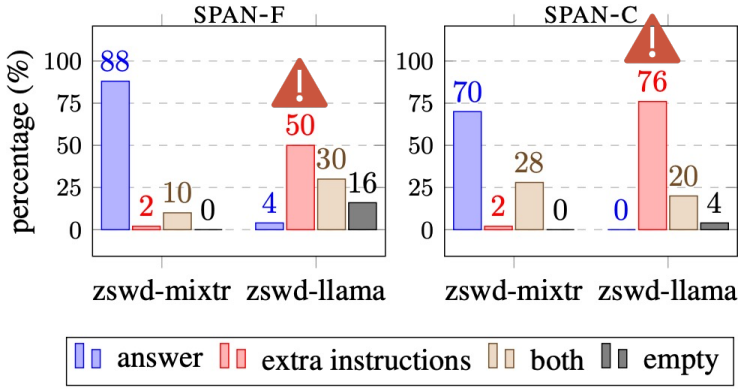
ZSWD-LLAMA
ZSWD-MIXTR

- We test LLAMA-3 8B & MIXTRAL 8X7B using prompts *with fallacy definitions* (WD)

Analysis of LLMs' *raw outputs* and *answers*

We conduct a **manual audit** of LLMs' outputs – *50 per model/setup, 400 in total*

- **Raw outputs**. Does the LLM provide an *answer*, *extra instructions*, *both*, or an *empty* response?
- **Actual answers**. Is the answer in the requested format (*format ok*)? Does the LLM provide extra *explain*ations, *wrong labels*, or repetitions (*repeat*)?

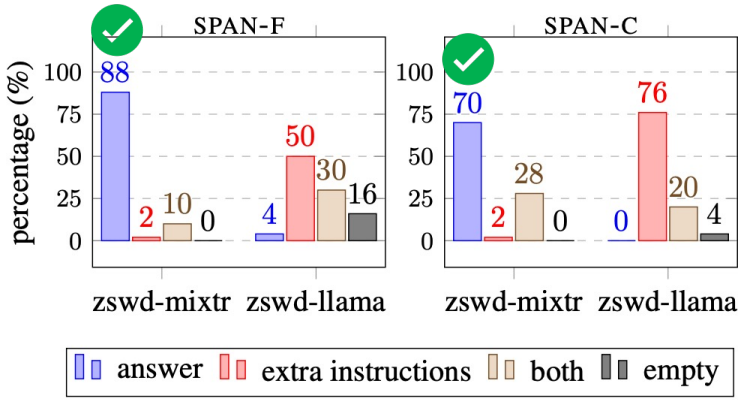


Check the paper *for results for* POST *tasks (similar findings)*

Analysis of LLMs' *raw outputs* and *answers*

We conduct a **manual audit** of LLMs' outputs – *50 per model/setup, 400 in total*

- **Raw outputs**. Does the LLM provide an *answer*, *extra instructions*, *both*, or an *empty* response?
- **Actual answers**. Is the answer in the requested format (*format ok*)? Does the LLM provide extra *explain*ations, *wrong labels*, or repetitions (*repeat*)?

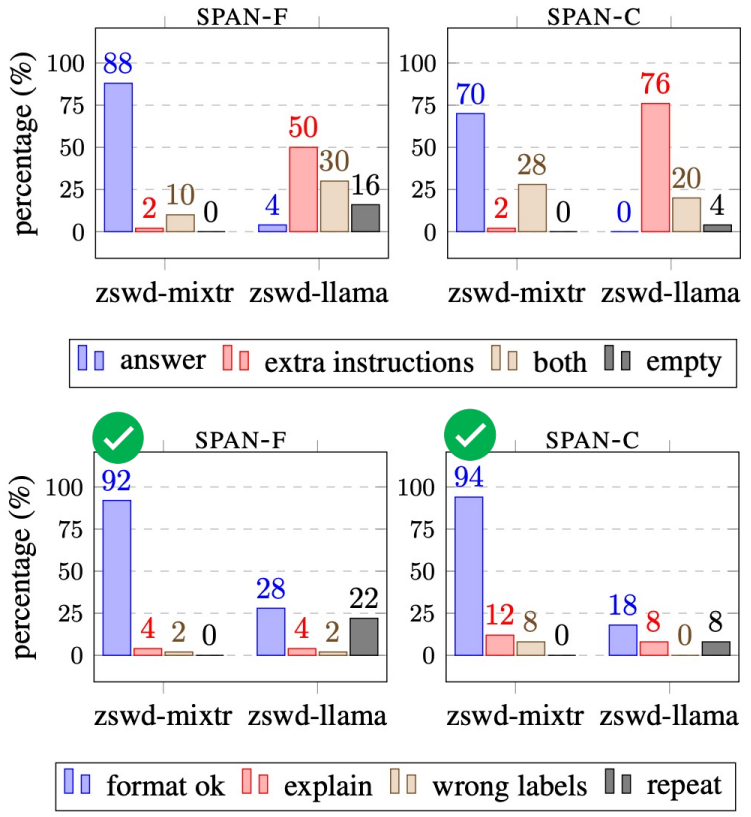


Check the paper *for results for* POST *tasks (similar findings)*

Analysis of LLMs' *raw outputs* and *answers*

We conduct a **manual audit** of LLMs' outputs – *50 per model/setup, 400 in total*

- **Raw outputs**. Does the LLM provide an *answer*, *extra instructions*, *both*, or an *empty* response?
- **Actual answers**. Is the answer in the requested format (*format ok*)? Does the LLM provide extra *explain*ations, *wrong labels*, or repetitions (*repeat*)?



Check the paper *for results for* POST *tasks (similar findings)*

Conclusions

We introduced 🐻 **FAINA**, the first fallacy detection dataset:

- Embracing **human label variation** at the **fine-grained** level of text segments
- Covering **multiple topics** and a **large time frame** to minimize topic/temporal biases
- Supported by detailed **insights on the annotation protocol** and **data collection**
- Accompained by an **evaluation framework**, **baselines**, and **insights on LLMs' outputs**

*We release 💾 **data**, 🐍 **code**, and 📕 **annotation guidelines** to foster research on fallacy detection & human label variation, and to support extensions to new languages, topics & annotators*