WorthIt: Check-worthiness Estimation of Italian Social Media Posts

Agnese Daffara[1],[2], Alan Ramponi[3] and Sara Tonelli[3]

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

CLiC-it 2025

What is “check-worthy”?

A claim is **check-worthy** and calls the attention of a **fact-checker** if:

- It is **factual** and **verifiable**, i.e., it presents an “assertion about the world that is checkable”[1].
- It is **not** “**easy** to fact-check by a layperson”[2].
- It is “likely to be **false**, is of **public interest**, and/or appears to be **harmful**”[2].

[1] Konstantinovskiy, O. et al. (2021), Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection, Digital Threats 2.

[2] P. Nakov et al. (2022), Overview of the CLEF-2022 Check-That! lab task 1 on identifying relevant claims in tweets, CLEF 2022.

What is “check-worthy”?

Check-worthiness estimation

(or *check-worthy claim detection*)

Why does it matter?

Check-worthiness estimation is an important step in the **fact-checking pipeline**, because it feeds to the fact checker only those posts that are **societally relevant** and **potentially impactful**, optimizing the verification process.

The task is well-known[3], but with some **limitations**:

- Datasets are mainly on **specific issues** (e.g. COVID-19) and a small **time period.**
- Existing datasets in Italian[4][5] contain only check-worthy claims to be **directly** fact-checked.
- The **relationship** between factuality and check-worthiness is not explored.

[3] E.g. in the shared task CheckThat!, organized by the CLEF initiative.
[4] Gili, L. et al. (2023), Check-IT!: A corpus of expert fact-checked claims for Italian, CLiC-it 2023.
[5] A. Scaiella, S. et al. (2024), Leveraging large language models for fact verification in Italian,CLiC-it 2024.

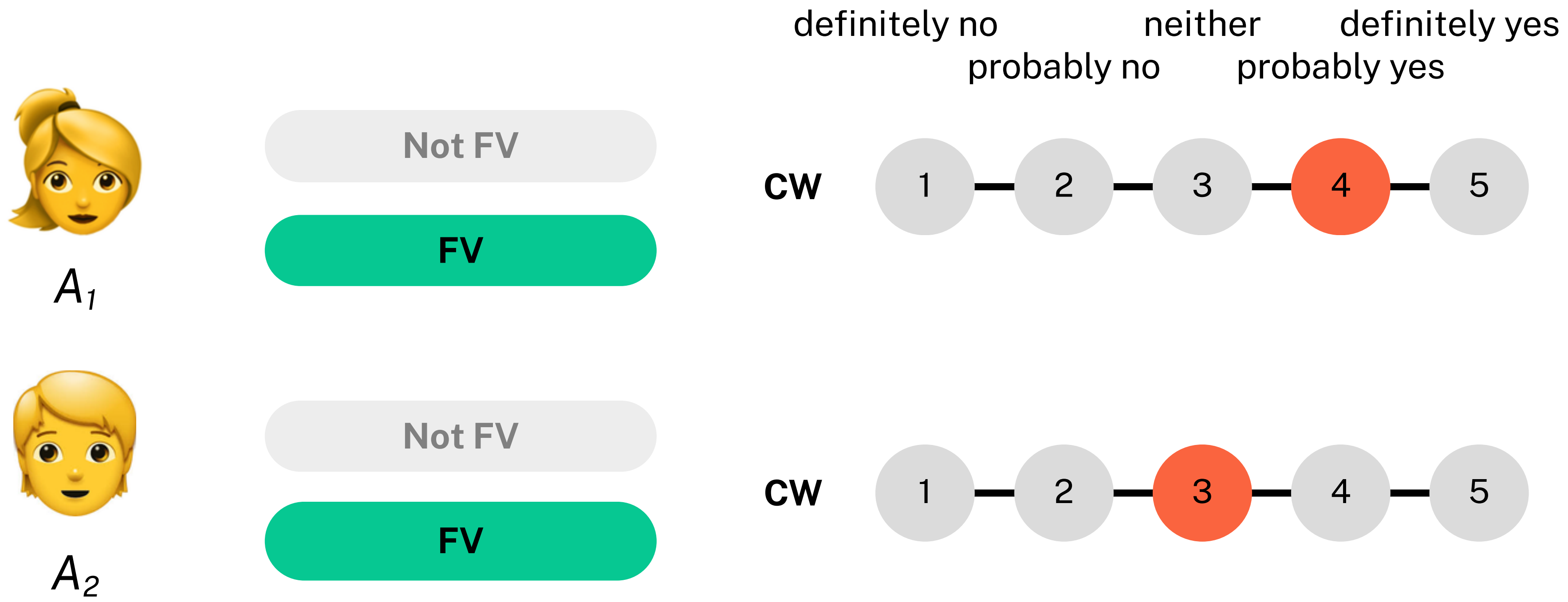
WorthIt dataset

- Dataset for **factuality/verifiability (FV)** and **check-worthiness (CW)** estimation.
- Focuses on Italian **social media posts**.
- Embraces **Human Label Variation (HLV)**.
- Two expert **annotators**.
- Four annotation **rounds** with discussion.

Example

"Tre ragazzi da Mali, Iraq e Mauritania, salvati stanotte a oltre 2000 mt. a Claviere, alta Valsusa. Migranti. Sotto la pioggia, con un principio di ipotermia."

EN: *"Three young men from Mali, Iraq, and Mauritania were rescued last night at over 2000 meters in Claviere, upper Valsusa. Migrants. In the rain, with the onset of hypothermia."*



WorthIt dataset

Data collection & sampling

2,160 post
83,315 tokens
38.6 avg token lenght

Public discourse on Twitter minimizing temporal & topic biases:

- **Multi-year**: ⏳ 6-year time frame (2017-01 — 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[6] for each month/topic.
- Resample posts by the same authors after their most impactful one.

[6] P. Nakov et al. (2022), Overview of the CLEF-2022 Check-That! lab task 1 on identifying relevant claims in tweets, CLEF 2022.

WorthIt dataset

Data collection & sampling

Partial overlap with **FAINA**[7]

[7] Alan Ramponi et al. (2025), Fine-grained Fallacy Detection with Human Label Variation, NAACL 2025.

WorthIt dataset

Data collection & sampling

Partial overlap with **FAINA**[7]

☐ In WORTHIT only ▨ In both WORTHIT and FAINA

Migration	120	120	120	120	120	120
Climate change	120	120	120	120	120	120
Public health	120	120	120	120	120	120
	2017	2018	2019	2020	2021	[illegible]

Check the poster in the next poster session!! :)

[7] Alan Ramponi et al. (2025), Fine-grained Fallacy Detection with Human Label Variation, NAACL 2025.

WorthIt dataset

Data annotation: statistics

The dataset is released with disaggregated labels to incentivate future studies on HLV.

Inter-annotator agreement (IAA) is calculated with Krippendorff's Alpha (α) after discussion, by keeping **natural disagreement**:

- **0.83** for factuality/verifiability.
- **0.69** for check-worthiness (lower as expected).

If a post is not factual/verifiable, annotators do not label it for check-worthiness.

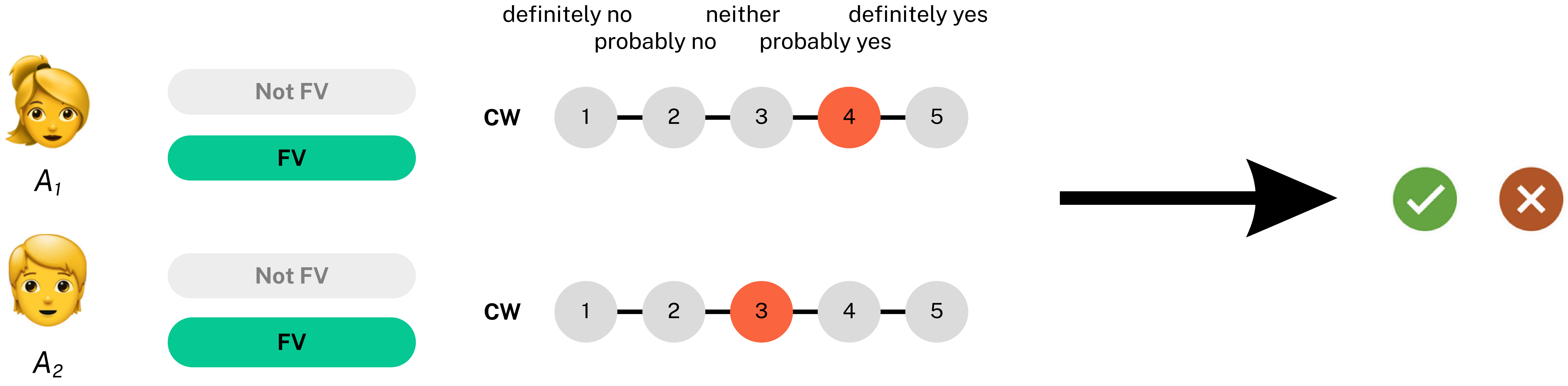
Annotator A_1					
	NO	YES			
FV	747 (34.6%)	1,413 (65.4%)			
CW		43 (2.0%)	342 (15.8%)	17 (0.8%)	807 (37.4%)	204 (9.4%)
		← NO			YES →

Annotator A_2						
	NO	YES				
FV	728 (33.7%)	1,432 (66.3%)				
CW		145 (6.7%)	380 (17.6%)	123 (5.7%)	574 (26.6%)	210 (9.7%)
		← NO				YES →

Experiments

Setup: label aggregation

- We **aggregate labels** for our experiment: a post is considered factual if both annotators agreed on ts factuality, and check-worthy if they gave positive labels (*probably yes* and *definitely yes*).

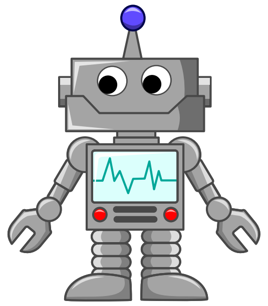


Experiments 

Setup: data splits

- **Data splits**: we divide WorthIt into k training and test sets using k**-fold cross-validation (**k **= 5)** preserving the label distribution. Training sets are further divided into development and train test:
  - 80% **training** and 20% **development** for encoder-based models.
  - 50% for retrieving few-shot **examples** and 50% as a **development** set for decoder-based models.

Experiments



Setup: models

Encoder-based models

Italian models:
- AlBERTo
- UmBERTo
- dbmdz's Italian BERT models:
  - BERT-it base
  - BERT-it xxl

Multilingual models:
- mBERT
- XLM-RoBERTa

Decoder-based models (instuction tuned)

Italian models:
- LlaMAntino-3-ANITA-8B
- Minerva-7B

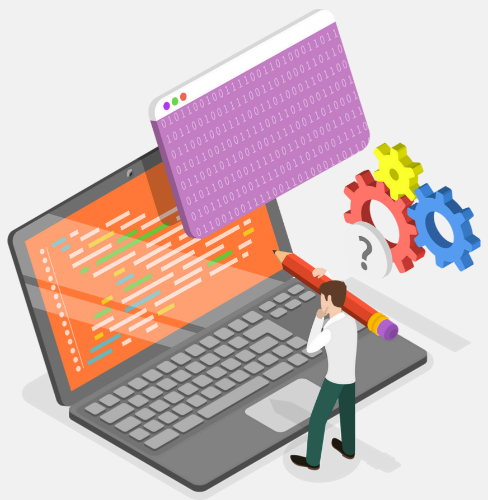
Multilingual models:
- Qwen2.5-7B
- Llama3.1-8B

For **fine-tuning**, we use the MaChAmp toolkit (v0.4.2)[8].

[8] R. van der Goot et al. (2021), Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP, EACL demos

Experiments

Setup: best prompt selection



Example set: we test the models over 5 sets of examples (5 examples each).

Language and guidelines: we test the models over 4 settings:

- **IT_NG**: Italian with guidelines
- **IN_G**: Italian without guidelines
- **EN_NG**: English without guidelines
- **EN_G**: English with guidelines

Final prompt configuration:

- Example set #1
- Without guidelines: **EN_NG**, **IT_NG**

Prompt for factuality/verifiability (en)

Classify the post as "factual" or "not factual". Answer only with "factual" or "not factual".
$FV_GUIDELINES
Examples:
$FV_EXAMPLES
Answer:
$POST_TEXT =

Prompt for factuality/verifiability (it)

Classifica il post come "fattuale" o "non fattuale". Rispondi solo con "fattuale" o "non fattuale".
$FV_GUIDELINES
Esempi:
$FV_EXAMPLES
Risposta:
$POST_TEXT =

Prompt for check-worthiness (en)

You classified the post as $FV_LABEL. Now classify the post as "check-worthy" or "not check-worthy". Answer only with "check-worthy" or "not check-worthy".
$CW_GUIDELINES
Examples:
$CW_EXAMPLES
Answer:
$POST_TEXT =

Prompt for check-worthiness (it)

Hai classificato il post come $FV_LABEL. Ora classifica il post come "check-worthy" o "non check-worthy". Rispondi solo con "check-worthy" o "non check-worthy".
$CW_GUIDELINES
Esempi:
$CW_EXAMPLES
Risposta:
$POST_TEXT =

Experiments

Setup: models configuration

Hypotesis: factuality/verifiability (FV) information can help predicting the check-worthiness (CW) of a post.

We test two configurations for each model

Encoder-based models:

- **SINGLE-TASK**: the model is fine-tuned with CW labels only.
- **MULTI-TASK**: FV serves as an auxiliary task.

Decoder-based models:

- **NOT-SEQUENTIAL**: the model is prompted directly for CW.
- **SEQUENTIAL**: the model is firstly instructed to classify the post based on FV, then the output label is incorporated into a prompt which instructs the model to assess CW of the same post.

Evaluation:

- Pos F1 (main metric)
- Pos Prec, Pos Rec and Acc.
- Mean average precision (mAP) for encoder-based models
- N. of “unknown” outputs for decoder-based models

Results

Encoder-based models

- FV as a support task **helps** improving the Pos F1 performance across all models.
- Best scores: **BERT-it xxl** in **MULTI-TASK** setting.

Model	Setting	Pos F_1
AlBERTo	SINGLE TASK	$0.7039_{\pm 0.03}$
	MULTI-TASK	$\underline{0.7107}_{\pm 0.02}$
UmBERTo	SINGLE TASK	$0.7247_{\pm 0.02}$
	MULTI-TASK	$\underline{0.7277}_{\pm 0.02}$
BERT-it base	SINGLE TASK	$0.7121_{\pm 0.02}$
	MULTI-TASK	$\underline{0.7146}_{\pm 0.03}$
BERT-it xxl	SINGLE TASK	$0.7332_{\pm 0.02}$
	MULTI-TASK	$\underline{\mathbf{0.7473}}_{\pm 0.02}$
mBERT	SINGLE TASK	$0.6767_{\pm 0.03}$
	MULTI-TASK	$\underline{0.6828}_{\pm 0.03}$
XLM-RoBERTa	SINGLE TASK	$0.7014_{\pm 0.02}$
	MULTI-TASK	$\underline{0.7138}_{\pm 0.02}$

Results

Decoder-based models

- FV **does not help** models predicting the check-worthiness (esp. true for multilingual models).

- Highest score: **LlaMAntino-3-ANITA-8B SEQ, EN**.
- **Minerva-7B** is the only model to produce “unknown” outputs.

Model	Setting	Lang	Pos F_1	Unknown
LlaMAntino-3-ANITA-8B	NOT SEQ	en	$0.6556_{\pm 0.03}$	0
		it	$0.6409_{\pm 0.02}$	0
	SEQ	en	$\underline{\mathbf{0.6771}}_{\pm 0.02}$	0
		it	$0.6111_{\pm 0.03}$	0
Minerva-7B	NOT SEQ	en	$0.3506_{\pm 0.01}$	$81_{\pm 2}$
		it	$0.3629_{\pm 0.01}$	$112_{\pm 8}$
	SEQ	en	$0.2944_{\pm 0.00}$	$127_{\pm 8}$
		it	$\underline{0.4442}_{\pm 0.02}$	$58_{\pm 4}$
Qwen2.5-7B	NOT SEQ	en	$0.5917_{\pm 0.02}$	0
		it	$\underline{0.6273}_{\pm 0.01}$	0
	SEQ	en	$0.5885_{\pm 0.01}$	0
		it	$0.6247_{\pm 0.02}$	0
Llama3.1-8B	NOT SEQ	en	$0.5470_{\pm 0.00}$	0
		it	$\underline{0.5616}_{\pm 0.01}$	0
	SEQ	en	$0.5585_{\pm 0.01}$	0
		it	$0.5584_{\pm 0.01}$	0

Discussion

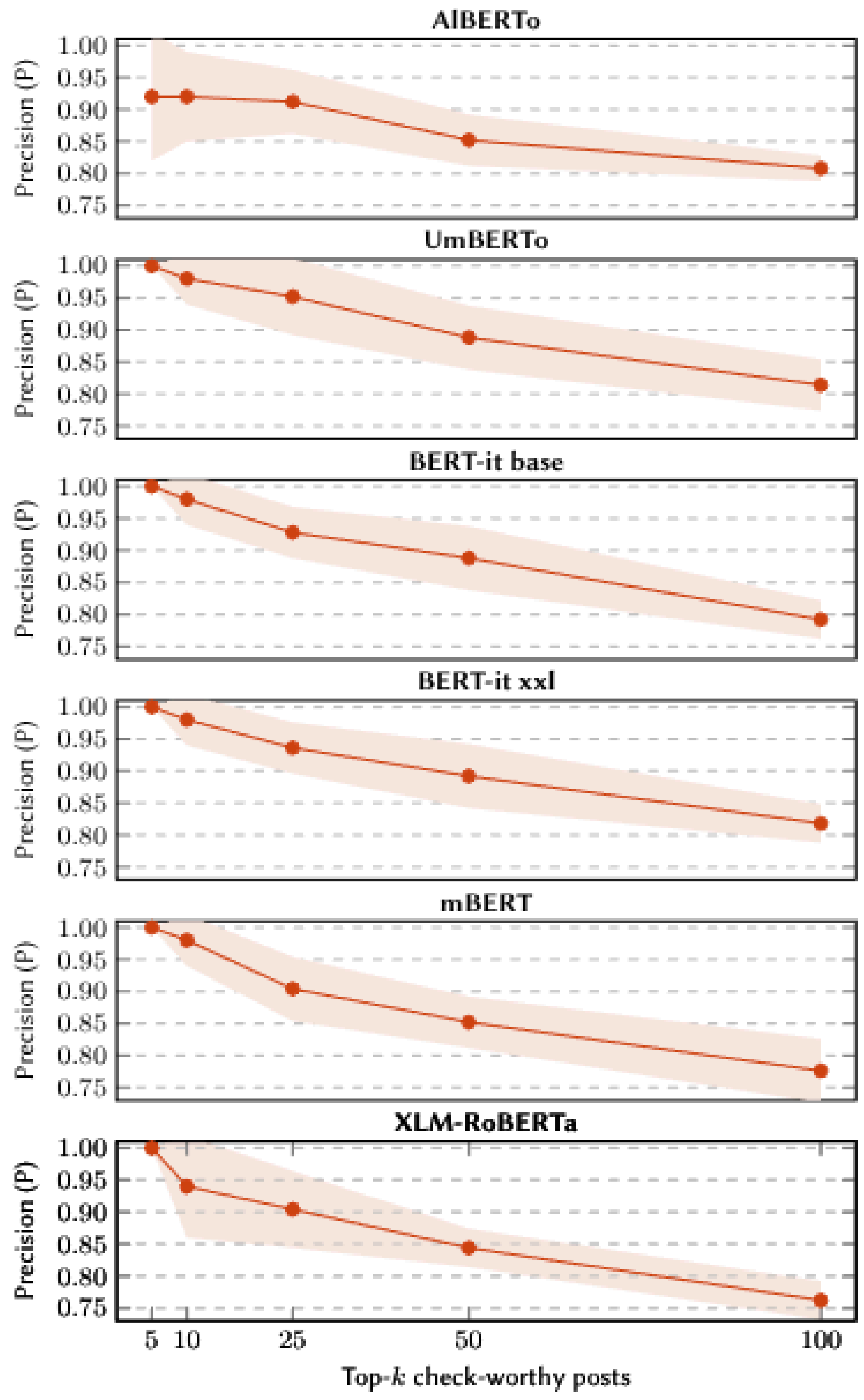
Ranking of posts by check-worthiness (encoder-based models)

Are encoder-based models good at ranking CW posts?

The ratio of posts correctly classified as check-worthy within the top-k recommended check-worthy posts (P@k) by all **encoder-based models** is:

- Precision is 0.90–0.95 at k = 25
- Precision is 0.80–0.85 at k = 100

→ These models **can help fact-checkers** in their daily routine!



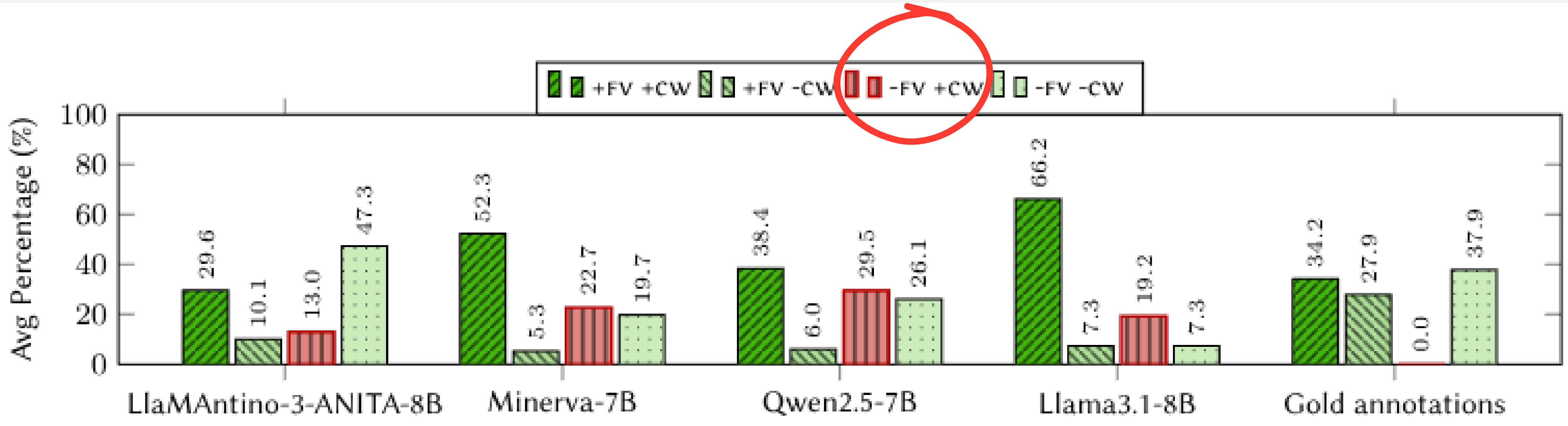
Discussion

Relation between FV and CW (decoder-based models)

Do decoder-based models understand the relation between FV and CW?

- Models tend to produce the invalid label combination **-FV +CW**.
- Models tend to avoid the combination **+FV -CW**, preferring to align the two labels rather than diversifying them.

→ These models seem to **not grasp** the relation between the two concepts



Conclusions

- We introduce **WorthIt**, the first dataset of Italian social media posts annotated for factuality/verifiability (FV) and check-worthiness (CW) that spans multiple years ⏳ and topics 🔄🌱🏥 while considering natural disagreement. This dataset partially overlaps with the dataset **FAINA** for fallacy detection.
- We conduct thorough check-worthiness estimation experiments with **encoder-** and **decoder-based models.**

Main finding

- **Encoder-based models** in a **multi-task** setting reach the best results → they can be used in fact-checking pipelines.
- **Decoder-based models fail to capture the relation** between FV and CW and produce **inconsistent** results → they require more caution.

GitHub repository:
https://github.com/dhfbk/worthit

Thank you!

Discussion

Correlation between models' outputs

What is the correlation between all models' outputs?

- We calculate the Pearson correlation coefficient (r) between all best models' predictions.
- Encoder-based models show strong positive mutual correlation ($r \geq 0.65$) $\longrightarrow$ **high consistency** in the predictions.
- Decoder-based models show low inter-model correlation → greater **output variability**.