Shrome at Touché: Soft-Vote Ensembling and Counter-Causal Augmentation for Causality Extraction

Touché at CLEF 2026 · Causality Extraction · Session 2

Shayan Sooratgar & Roham Zendehdel Nobari

Department of Computational Linguistics, University of Zurich

Universität Zürich[UZH]

Same verb, opposite meaning

■ cause ■ effect

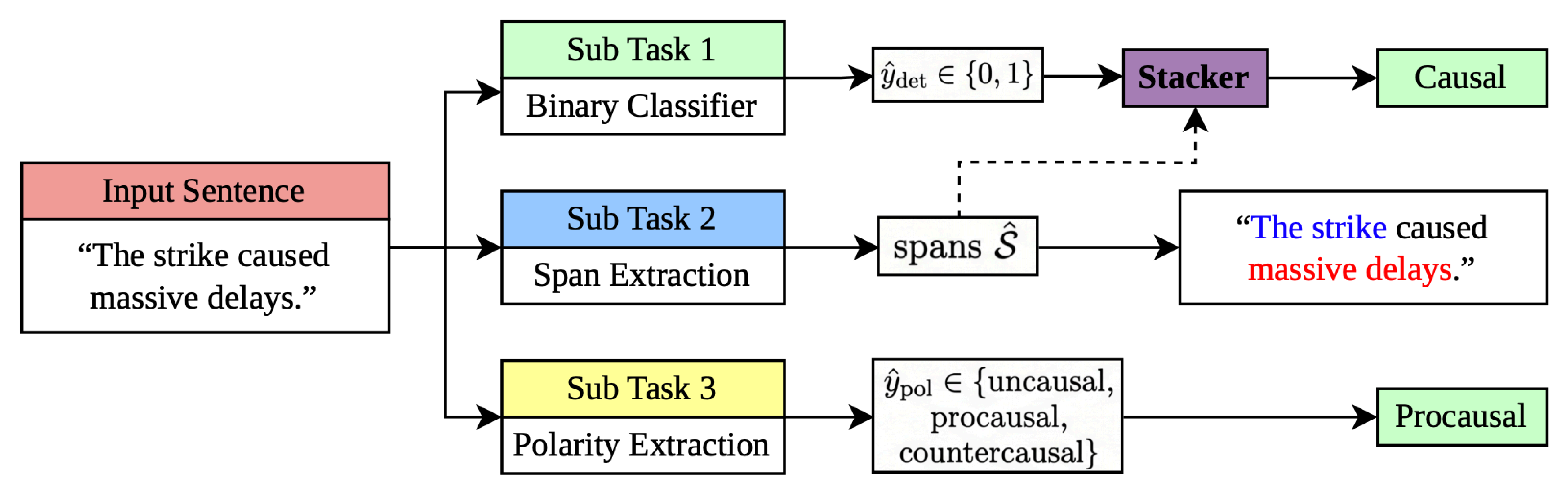
procausal The strike caused massive delays.

counter-causal **It is falsely believed that** the strike caused massive delays.

1. **Detection**
   Is it causal?
2. **Extraction**
   Where are cause and effect?
3. **Polarity**
   Pro-, counter- or uncausal?

> Counter-causal is the **smallest** polarity class (886 training rows) and the most **negation-heavy** (67.8% vs 15.8% for procausal).

One model per subtask, one rule connecting them



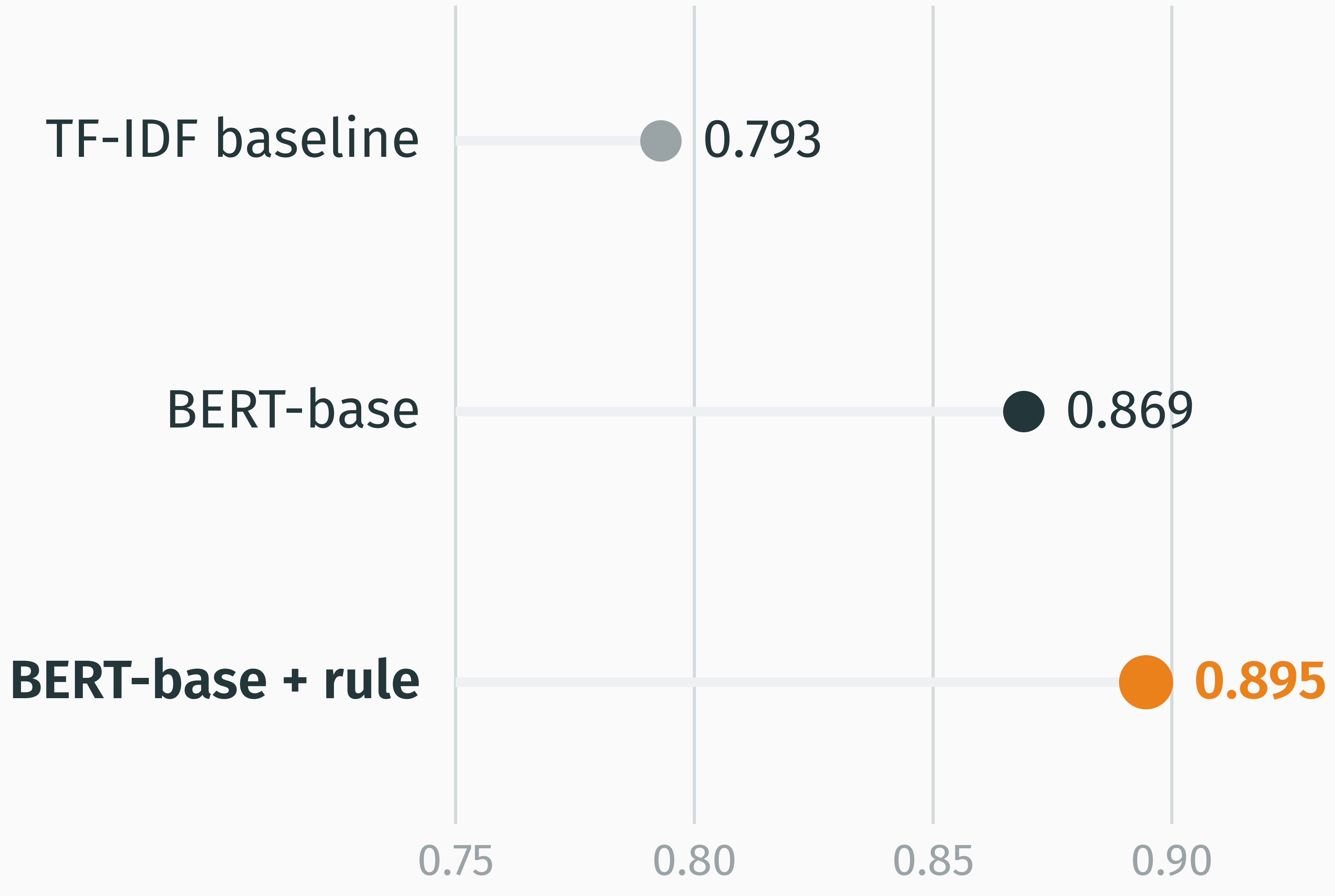
The one cross-task rule: causal = detector says causal **and** extraction finds at least one span.

Our contributions: **soft-voting** in extraction, **data augmentation** in polarity.

Figure 1 of the paper

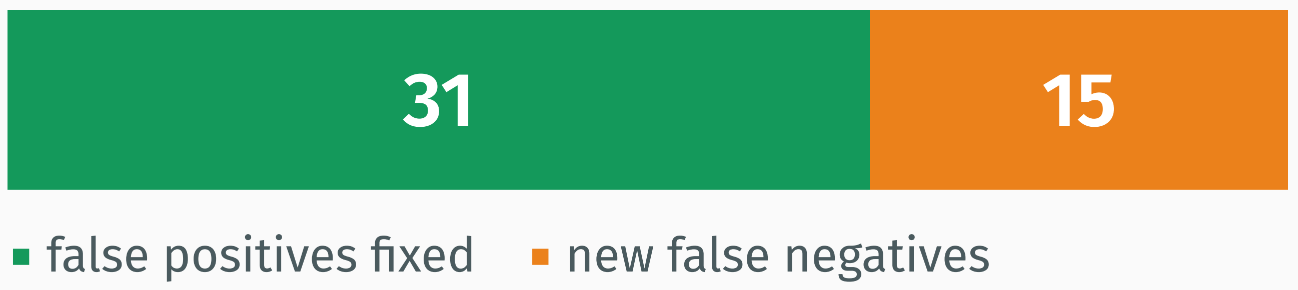
BERT, plus a check against extraction

Detection F1 on dev (causal class)

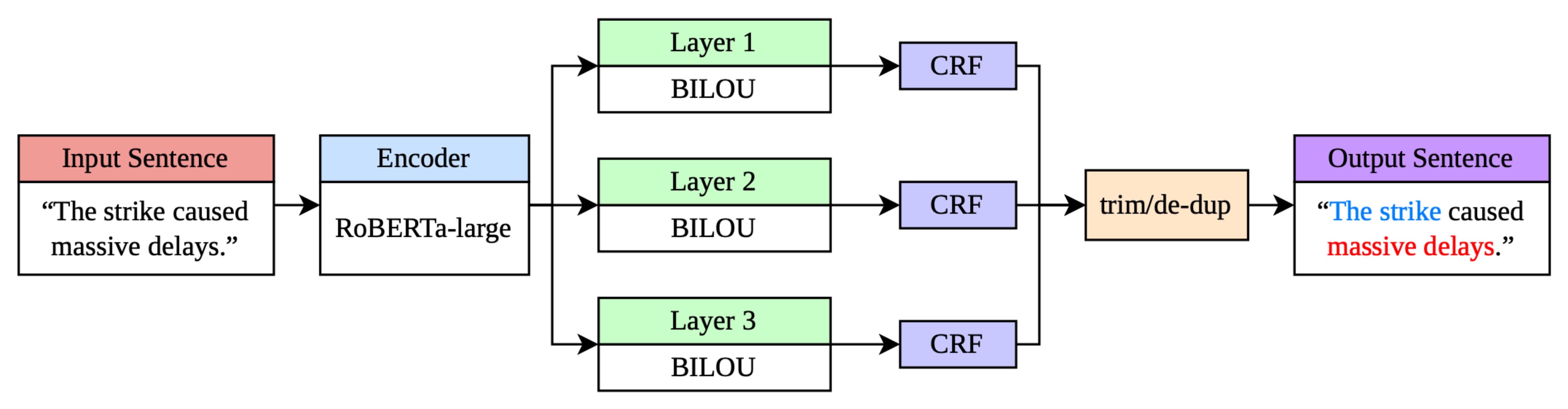


No span found? Then it is probably not causal.

Flips on dev



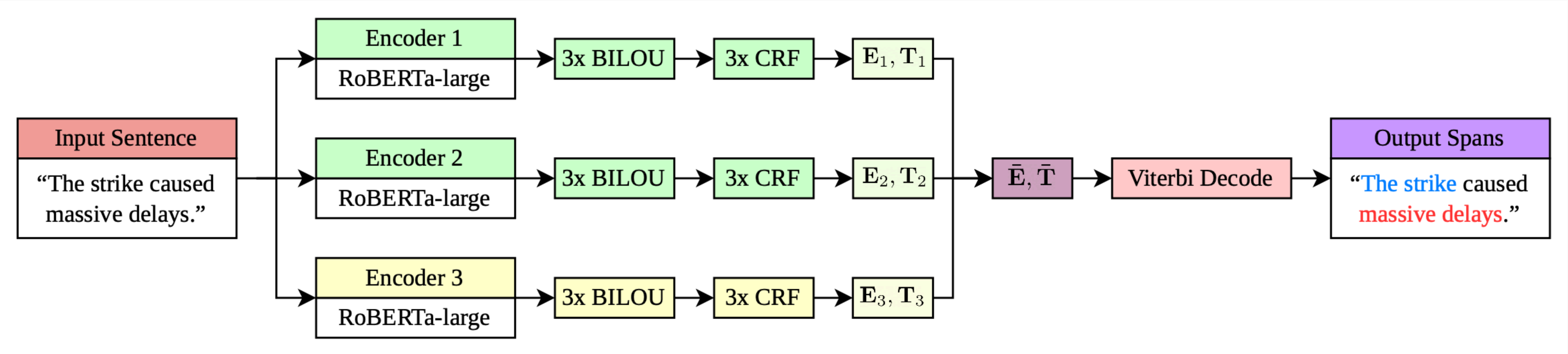
Three tagging layers, one CRF each



Three layers so one sentence can hold several overlapping cause/effect spans.

Best single-model exact-match F1 on dev

Don't vote on spans. Average, then decode.



Usual: hard vote

1. Each model decodes its own spans
2. Majority vote over the spans
3. Near-misses are thrown away

Ours: soft vote

1. Average emission + transition scores
2. One Viterbi decode on the average
3. Start scores 0.8 · 0.6 · 0.7 → **0.7**

Same 3 models	Exact F1
Hard vote	0.578
Soft vote	0.589
Final panel	**0.593**

Figure 4 of the paper · +1.07 points from changing only the combiner · single model: 0.574

Teach the model how causation is denied

Nine patterns of denying causation

direct negation	negated context	lack of counterfactuality
lack of effect	implicit lack of effect	negative causation
usual inverse effect	**coincidence**	**temporal order**

Pattern examples

“It is **not** that X caused Y.”

“X and Y happened **coincidentally**.”

“He is happy. He did **not** win the lottery.”

GPT-5.4 rewrites a training sentence

one pattern per prompt, cause and effect kept unchanged

↓

Structural check

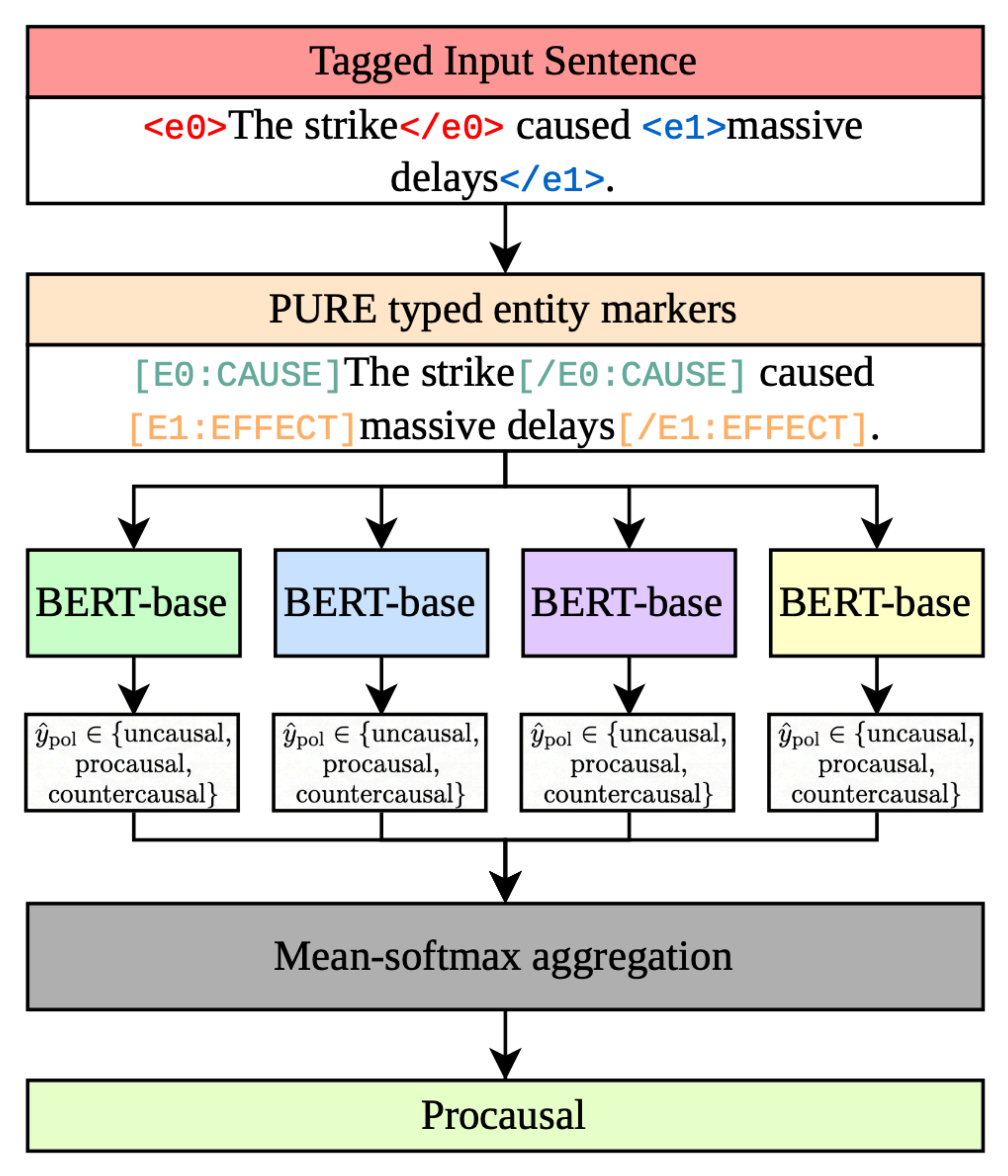
93.2% pass (structure, not meaning)

Counter-causal training rows

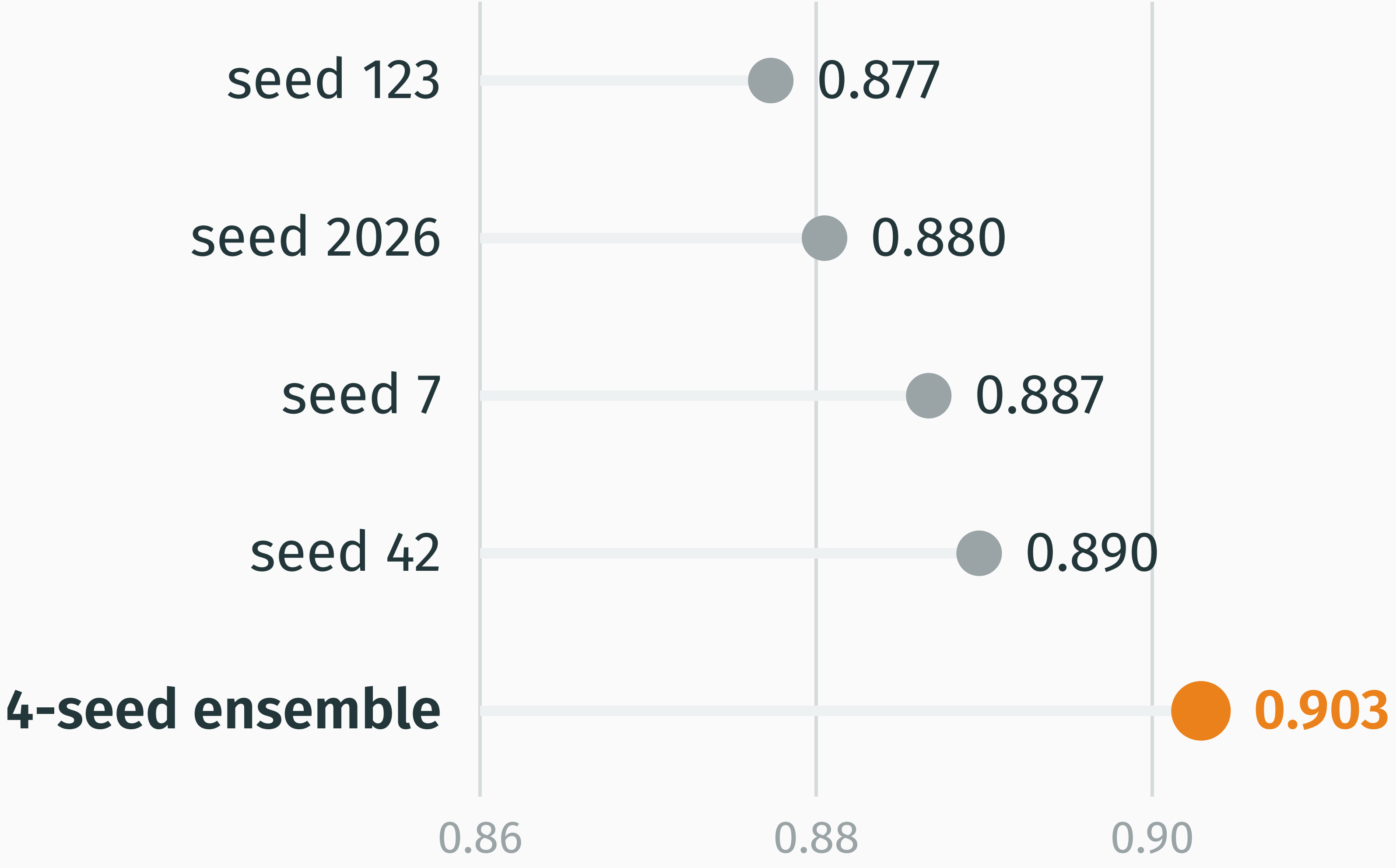
Dev macro F1 (single seed)

Patterns adapted from Hagen et al., 2025 · checks confirm structure, not meaning

Mark the spans, average four models



Dev macro F1



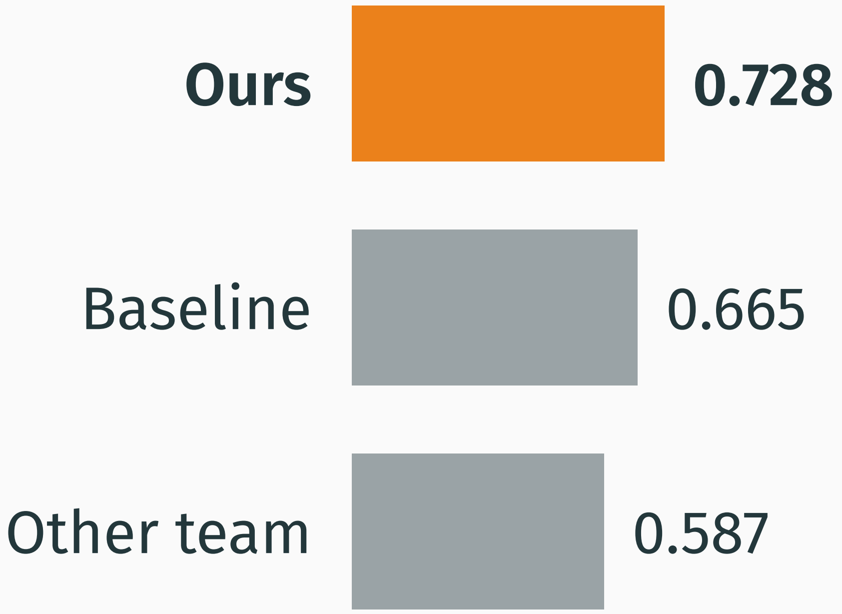
Counter-causal F1 of the ensemble: **0.891**

> **Biggest remaining error:** uncausal predicted as causal (21 of 264). Pro- vs counter-causal confusion: 5–6 per direction.

Figure 5 of the paper · seed-42 confusion matrix · CCNC dev split

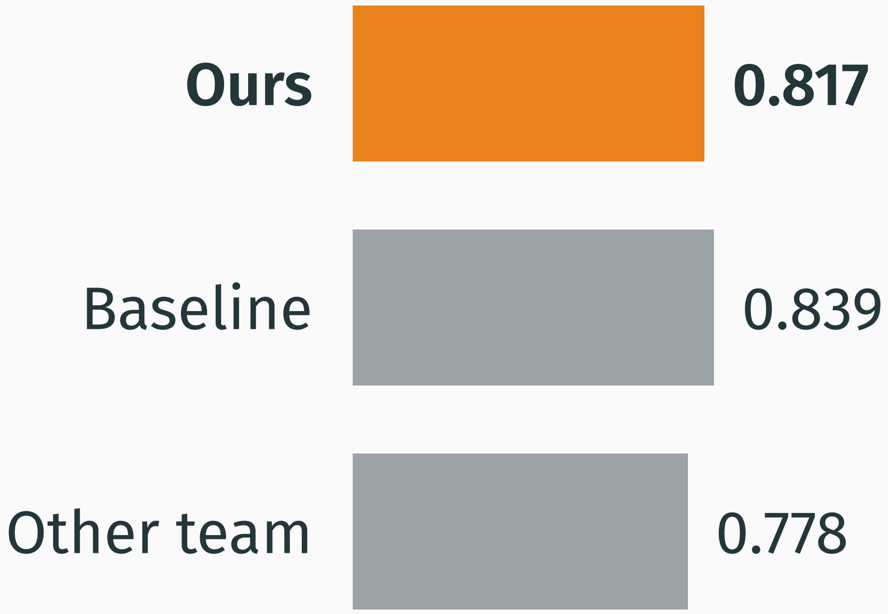
Official results

Extraction · granularity-adj. F1*



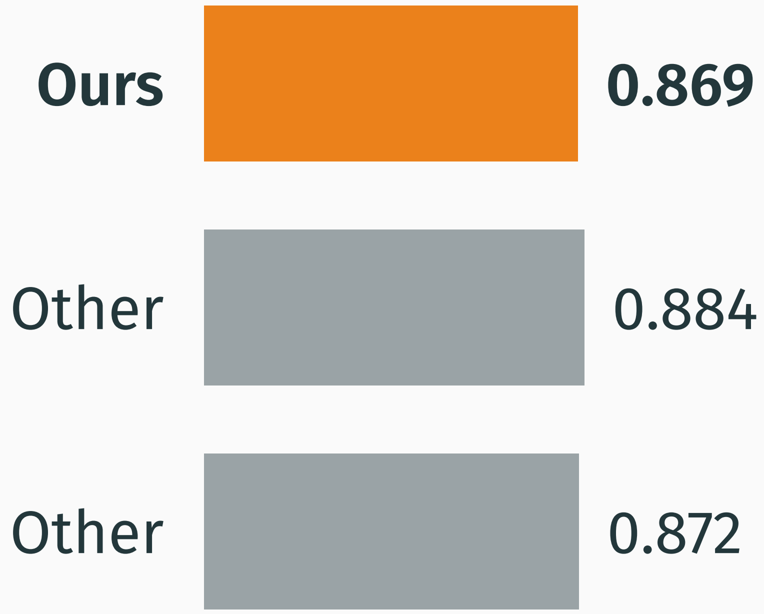
highest overall

Polarity · macro F1



best team

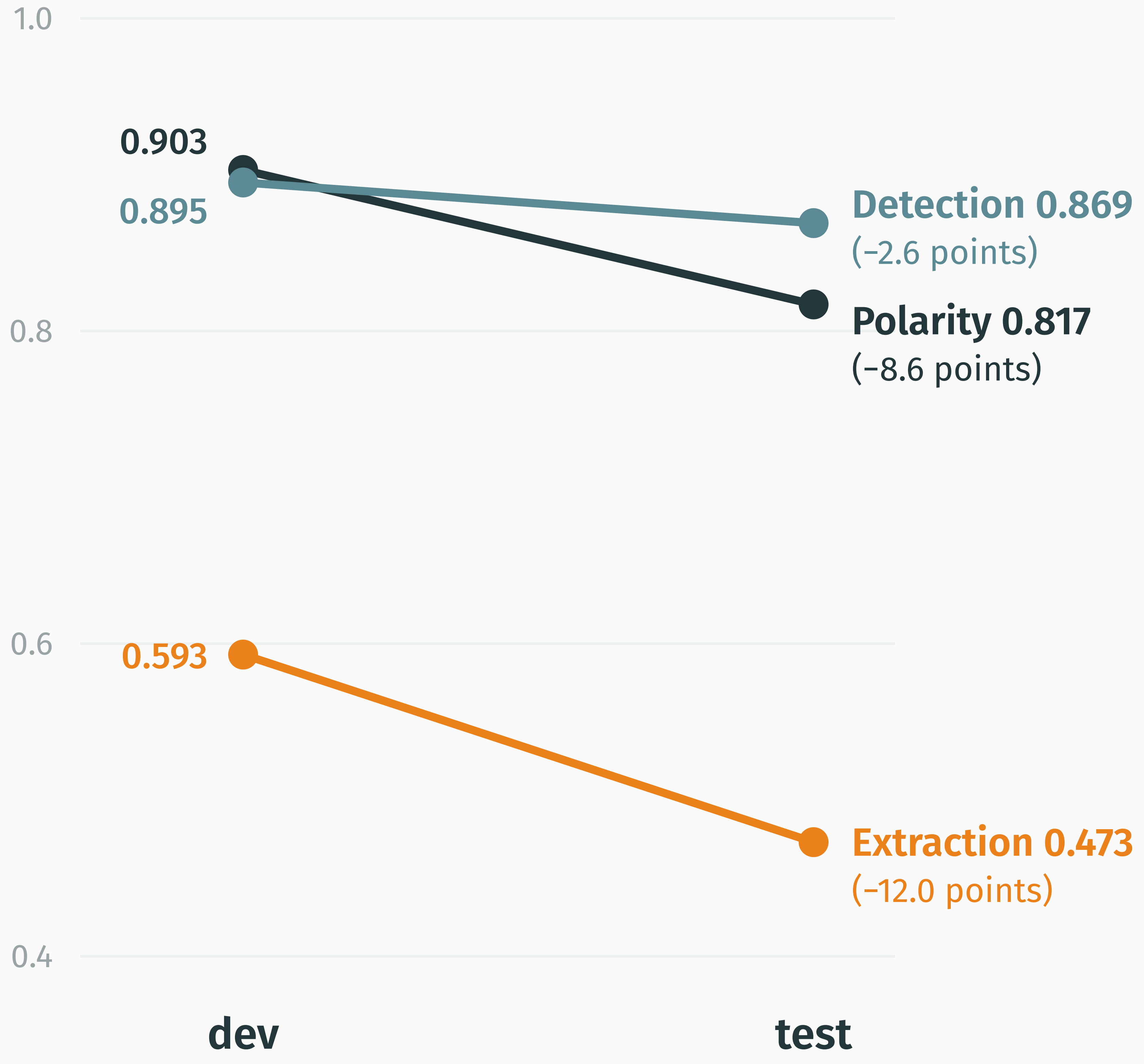
Detection · F1



third, highest recall

*Organizers’ final causal-only evaluation of extraction.

Dev was optimistic



Why

Panel, rule and ensemble size were all chosen on the same dev split.

Next

Negation-scope feature for polarity.
New encoders and span architectures for extraction.

Dev: CCNC dev split · Test: official TIRA evaluation (extraction: all-samples exact-match F1)

Takeaways

1. **Average before you decode.**
2. **Teach the model how causation is denied.**
3. **Check dev gains on test.**

Code, prompts and generated data

github.com/rzninvo/NegaCausal

Questions? shayan.sooratgar@uzh.ch · roham.zendehdelnobari@uzh.ch

Backup: a post-hoc NLI filter barely moves the needle

> Hypothesis "This sentence describes a causal event involving '**span**'." → DeBERTa-v3-large NLI → keep the span if entailment ≥ 0.1

Stage (dev)	Exact F1	Overlap F1
Soft-vote panel	0.5933	0.7800
+ NLI verifier	**0.5953**	**0.7781**

Removes 3 spans, all truncated fragments of gold spans. Within noise. A cause→effect pair template lost 5.8 points.

Backup: training setup

	Detection	Extraction (per seed)	Polarity (per seed)
Encoder	BERT-base	RoBERTa-large	BERT-base
Learning rate	2e-5	5e-5 (CRF 3e-4)	2e-5
Batch · max length	32 · 128	16 · 256	32 · 256
Epochs · early stop	5 · none	≤30 · overlap F1	≤12 · macro F1
Seeds	42	123, 2026, 42 (TAPT)	42, 123, 2026, 7
Extra data	—	CNCv2 (+UniCausal for TAPT seed), EDA	+2,433 generated counter-causal

Single RTX 5090 · all seven deployed models train in under 12 GPU-hours · dev inference under 60 s.