arginvariant at Touché:
Error-Driven Prompt Refinement for Argument Identification

Touché at CLEF 2026 Jena, Germany 22 September 2026

Midhun Kanadan

Maximilian Heinrich

Benno Stein

Bauhaus-Universität Weimar

webis.de

The Problem: Guidelines Define What Counts as an Argument

The same 340 tweets, re-annotated under three different guidelines

"IMHO Brexit was all about democracy?"

Corpus	Guideline it applies	Gold label
TACO	Twitter: inference vs. information	✓ Argument
TAPE	Persuasive essays (Stab & Gurevych)	✗ No-Argument
TAUS	US presidential debates (Haddadan et al.)	✓ Argument

119 of 340 test tweets (35%) get different labels under the three guidelines. Being an argument is a property of the guideline, not of the sentence.

- Test item 88, shown without its two leading @-mentions. Feger et al. (ACL 2025) show this for fine-tuned **encoder** models: 97% of cross-dataset experiments fall below the in-dataset mean.

Touché 2026: Generalizability of Argument Identification in Context

Two properties of the task that shape our approach

1. The guideline is part of the input

- **Input:** sentence + source document + corpus annotation guideline
- **Output:** *Argument* or *No-Argument*

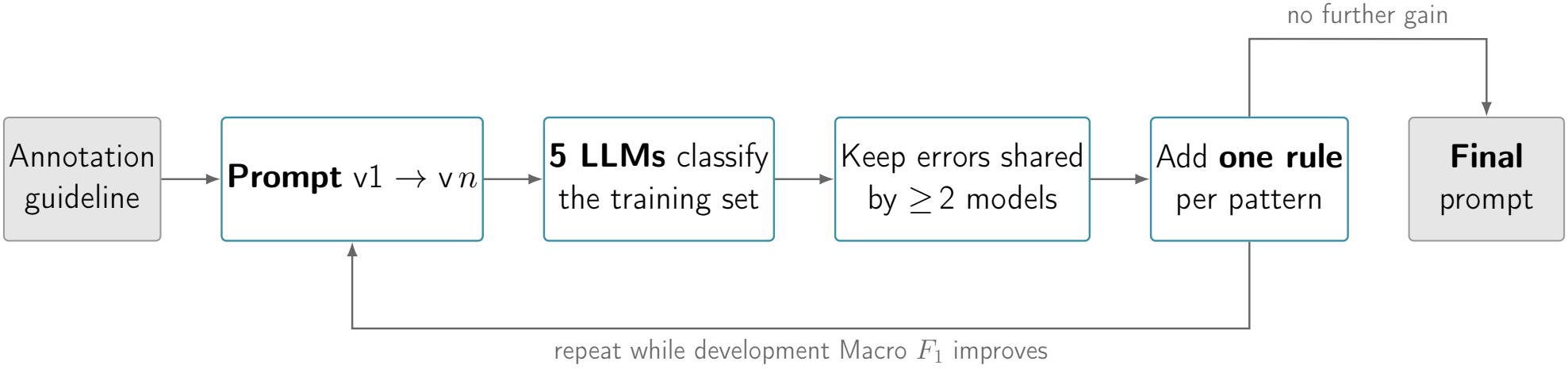
2. The primary evaluation has no labels

- 340 tweets, each labeled under TACO, TAPE and TAUS
- No training or development labels for any of them
- A guideline-blind system gives all three *the same* label

Ranking metric: unweighted mean Macro F_1 over the three. The ten benchmark corpora, which do have labels, form the *supplementary* metric.

arginvariant: Error-Driven Prompt Refinement

Distill each corpus's annotation guideline into a prompt and let a large language model apply it at inference time. Zero parameter updates.



Key design decision: only errors shared by at least two of the five LLMs drive a prompt update – a mistake only one model makes says more about that model than about the prompt.

What the Refinement Loop Actually Learns

AEC corpus: the largest gain (+0.558 Macro F_1)

Initial prompt (v1) Macro F_1 = **0.419**

"Does the sentence make a substantive point or take a side?"

$\longrightarrow$

Refined prompt (v5) Macro F_1 = **0.977**

"Classify by discourse-connective opener alone (so, but, if, first, however, . . .): present $\rightarrow$ Argument, absent $\rightarrow$ No-Argument, regardless of content."

146 of the 340 development sentences were false positives shared by two models – and *none* of them opened with a discourse connective.

- v1 tested *meaning* – the gold labels track the connective that opens the sentence.
- The loop recovered the corpus's own design: the same four cues its authors sampled on – so, but, if, first.

The rule is itself a surface shortcut – but now it is explicit and inspectable, not hidden in model weights.

Effect of Prompt Refinement

Development set, Macro F_1 – same model before and after, so the gain is the prompt's

Corpus	Model	v1	Best	Version	Δ
ABSTRCT	gemma3:27b	0.728	0.888	v5	+0.160
ACQUA	gemma4:26b	0.650	0.835	v5	+0.185
AEC	gemma4:26b	0.419	0.977	v5	**+0.558**
AFS	gemma3:27b	0.553	0.800	v4	+0.247
ARGUMINSCI	gemma4:26b	0.660	0.784	v3	+0.124
FINARG	gemma2:27b	0.612	0.735	v6	+0.123
IAM	deepseek-r1:14b	0.678	0.779	v5	+0.101
PE	gemma3:27b	0.581	0.728	v6	+0.147
SCIARK	gemma3:27b	0.615	0.800	v6	+0.185
USELEC	mistral:7b	0.548	0.692	v5	+0.144
Mean	—	0.604	0.802	—	**+0.198**

Refinement improves every one of the ten corpora, by +0.198 on average.

Results on the Blind Test Set

Official leaderboard – Main metric, best run per team

Rank	Team (run)	TACO	TAPE	TAUS	Main	Suppl.
1	`context-awakens`	0.8408	0.7604	0.7853	0.7955	0.7308
2	`arginvariant` (Run 1)	0.8133	**0.7757**	0.7612	0.7834	0.7899
3	`the-wildcards` (hybrid)	0.8265	0.7465	0.7735	0.7822	0.8128

- **2nd place** – and the best TAPE score of all submissions

The same prompts, on two models we never refined on

Model	v1	Refined	Δ
GPT-5 nano	0.6928	0.7461	+0.0533
Gemini 3.5 Flash	0.7785	**0.7945**	+0.0160

- Gemini: 0.10 pp below the winner – the gap is model capacity

Conclusions and Future Work

- Argument identification is a guideline-following problem.
- Written guidelines are incomplete – refinement recovers the conventions they omit.
- Shared-error aggregation makes the prompts model-agnostic; refinement adds +0.198 mean Macro F_1.

github.com/Midhun-Kanadan/

arginvariant-touche2026

Next

Automate the loop: an LLM agent distills the guideline into the first prompt and turns a reviewed error pattern into the rule – the two steps that are still manual.

Thank you!

Backup: AEC – the Evidence Behind the Rule

Why a surface rule, and not a better semantic definition

v1 Macro F_1 = **0.419**

"Does the sentence make a substantive point or take a side?"

146 of 340 development sentences were false positives shared by two models.

Not one opened with a discourse connective.

"Sadly, creationism is an assertion of dogmatism."
gold No-Argument, v1 said Argument

v5 Macro F_1 = **0.977**

"Classify by discourse-connective opener alone: present → Argument, absent → No-Argument, regardless of content."

All 8 shared false negatives *did* open with one.

"So what do you end up with?"
gold Argument, almost no semantic content

Swanson et al. (2015) sample AEC on exactly these cues – their IMPLICIT MARKUP hypothesis.

Holds on unseen data: 0.959 test Macro F_1. AEC is the *exception* – AFS and ACQUA were fixed by narrowing an over-broad definition.

Backup: the Three Submitted Runs

All three use the same refined prompts and differ only in model assignment

Run	TACO	TAPE	TAUS	**Main**	**Suppl.**
Run 1 – per-corpus model	0.8133	0.7757	0.7612	**0.7834**	**0.7899**
Run 2 – `gemma2:27b` uniform	0.8133	0.7757	0.7598	0.7829	0.7533
Run 3 – `gemma3:27b` uniform	0.8101	0.7204	0.7377	0.7561	0.6531
`context-awakens` (1st)	0.8408	0.7604	0.7853	0.7955	0.7308

- Run 2 matches Run 1 on Main (0.05 pp) but loses 3.7 pp on Supplementary – per-corpus model selection helps in-domain, not the primary metric.
- Run 3 ranked sixth; its weak spot is TAPE (0.7204), so `gemma3:27b` applies the essay guideline to tweets less well.

Backup: Models and Per-Corpus Assignment

Five open-source models, served with Ollama at temperature 0

Refinement pool

- `deepseek-r1:14b`
- `gemma2:27b`
- `gemma3:27b`
- `gemma4:26b`
- `mistral:7b`

Run 1 assignment

Corpus	Model
ABSTRCT, AFS, PE, SCIARK	`gemma3:27b`
ACQUA, AEC, ARGUMINSCI	`gemma4:26b`
FINARG	`gemma2:27b`
IAM	`deepseek-r1:14b`
USELEC	`mistral:7b`
TACO, TAPE, TAUS	`gemma2:27b`

No model dominates: `gemma3:27b` wins four corpora, `gemma4:26b` three.

Benchmarks routed by development Macro F_1; the Twitter variants had no labels, so PE and USELEC served as proxies.

Backup: the 119 Cross-Guideline Disagreements

Where the three guidelines demand different labels for the same tweet

TACO / TAPE / TAUS	Tweets
A / N / N	67
A / N / A	17
N / A / A	14
N / N / A	9
N / A / N	8
A / A / N	4
Total	**119**

A = Argument, N = No-Argument. The other 221 tweets get the same label everywhere.

Questions are the clearest single cause

Of the 30 test tweets ending in a question mark:

TACO	**23**	rhetorical questions are Inference
TAUS	**10**	questions can contain claims (rule 5)
TAPE	**0**	no rule for questions at all

24 of those 30 are among the 119 – one unwritten difference explains a fifth of all disagreements.