The Wildcards at Touché: Guideline-Conditioned Argument Identification with Dataset-Level Expert Routing

Touché at CLEF 2026

Can a model see which guideline rule is active?

Muhammad Qasim Hussain · Maximilian Heinrich · Midhun Kanadan · Benno Stein

Bauhaus-Universität Weimar

One tweet can have different labels

An example Tweet from the TACO dataset: “@Carol_Tucker @Sean_Moore Genius.”

TACO
original Twitter rule

ARGUMENT

TAPE
PE-based rule

NO-ARGUMENT

TAUS
USELEC-based rule

NO-ARGUMENT

MAIN

340 tweets × 3 rules
= 1,020 scored decisions

SUPPLEMENTARY

10 in-domain datasets
one definition per dataset

LATER DIAGNOSTIC

For one tweet, are all three rule-specific decisions correct?

The core idea: route each dataset to the classifier chosen for its rule

Before the blind test, each dataset is assigned one model-input pair; that route is then fixed.

1. READ THE TASK

Dataset
+
its annotation rule

→

2. LOOK UP THE ROUTE

Frozen map

one model-input pair
for the whole dataset

→

3. USE THE CHOSEN CLASSIFIER

Chosen classifier

labels every sentence
Argument or No-Argument

How we built the fixed routing map

We compared candidate classifiers and inputs, then used the evidence available before the blind test.

1. TEST CANDIDATE CLASSIFIERS

Fine-tuned medium and small sized language models

Prompted language models

→

2. COMPARE TWO INPUTS

Sentence only

versus

Guideline + sentence

→

3. USE AVAILABLE EVIDENCE

IN-DOMAIN (10 datasets)
Labels available
Highest development macro-F1

TWITTER (3 datasets)
Labels hidden
Strongest guideline reaction with plausible balance

Selected experts per dataset

In-domain: development macro-F1 · Twitter: predicted-Argument-rate spread

Selected expert	Datasets
Mistral-7B FT + guideline	ABSTRCT, AEC, AFS, FINARG, IAM, SCIARK
Llama-3.1-8B FT + guideline	USELEC
Llama-3.1-8B FT — sentence only	**ARGUMINSCI, PE**
DeepSeek V4 Pro + guideline	ACQUA

Twitter (TACO, TAPE, TAUS): DeepSeek V4 Pro + guideline · selected because its Argument prediction rate changed by 15.3 percentage points across the three guidelines.

3 Submissions: one fixed model and two routed systems

Scores: Main = mean macro-F1 over 3 Twitter definitions · Supplementary = mean macro-F1 over 10 in-domain datasets
Argument recall = recall of the Argument class over 1,020 Twitter decisions

Submission	What it uses	Twitter classifier	Main	Supplementary	Argument recall
Solo	One model for every dataset	Mistral-7B	0.594	0.815	**0.270**
Hybrid	Development winners in-domain; spread-selected model on Twitter	DeepSeek V4 Pro	0.782	0.813	**0.807**
Local	Same routes, self-hosted models	DeepSeek-R1-Qwen-14B	0.769	0.806	**0.780**

All three receive the active guideline on Twitter. Routing switches the classifier and repairs a recall failure.

Better overall classification is not precise guideline use

Tweet group	Tweets	Solo: all 3 correct	Hybrid: all 3 correct	Local: all 3 correct
All guidelines require the same label	221	159 (72%)	184 (83%)	185 (84%)
Guidelines require different labels	119	0 (0%)	10 (8%)	11 (9%)

RECOVERED LABEL SHIFT

“Genius.”

Gold A / N / N
Hybrid A / N / N

FAILED LABEL SHIFT

“The sun never sets on the British empire #Brexit”

Gold A / N / N
Hybrid A / A / A

A = Argument · N = No-Argument · order = TACO / TAPE / TAUS · not an official ranking metric

We next tested whether information beyond the sentence could help.

Could information beyond the sentence help?

FOUR STUDIED DATASETS

ABSTRCT · clinical-trial abstracts

PE · persuasive student essays

SCIARK · multidisciplinary scientific abstracts

FINARG · earnings-call questions and answers

340 released test sentences per dataset

TWO WAYS TO ADD INFORMATION

TARGET SENTENCE

+ original passage

OR

+ generated abstraction of the passage

summary · argument roles · preserved features · links to source sentences

Same classifier family → Argument / No-Argument

Same targets, labels, splits, and seeds across the comparison.

The follow-up improved four studied datasets but did not beat Main

Dataset	Submitted Hybrid	Later system
ABSTRCT	0.900	**0.918**
PE	0.771	**0.844**
SCIARK	0.823	**0.865**
FINARG	0.673	**0.718**

SUPPLEMENTARY

Best submitted (Solo) 0.815
Follow up study combination 0.831
Other six keep Hybrid predictions.
Not submitted or officially ranked.

MAIN

Official Hybrid 0.782
Later passage + guideline 0.776
The official score remained higher.

The follow-up also changed the training setup, so the higher in-domain scores cannot be credited to context alone.

Higher scores did not mean reliable definition following

01 Guidelines help selectively
The same added rule helped some model families and hurt others.

02 Routing improved the official system
One frozen expert per dataset raised Main from 0.594 to 0.782.

03 **Exact definition following remained rare**
Only 10 of 119 required label changes were fully recovered.

The full paired comparison is not monotonic in model size

System	Sentence	Guideline	Change
Mistral-7B FT	.760	.815	+5.6 pp
Llama-3.1-8B FT	.760	.800	+4.0 pp
DeepSeek-R1-Qwen-14B FT	.703	.736	+3.3 pp
TinyLlama-1.1B FT	.730	.684	−4.6 pp
SmolLM2-1.7B FT	.656	.636	−2.0 pp
Qwen2.5-Coder-1.5B FT	.561	.625	+6.4 pp
DeepSeek-R1-Qwen-1.5B FT	.496	.603	+10.7 pp
Qwen2.5-0.5B FT	.546	.595	+4.9 pp
DeepSeek V4 Pro prompted	.659	.685	+2.6 pp
DeepSeek-R1-Qwen-14B prompted	.655	.682	+2.7 pp

The full route mixes guideline-conditioned and sentence-only experts

Dataset(s)	Hybrid routed expert
ABSTRCT, AEC, AFS, FINARG, IAM, SCIARK	Mistral-7B FT + guideline
USELEC	Llama-3.1-8B FT + guideline
ARGUMINSCI, PE	**Llama-3.1-8B FT — sentence only**
ACQUA	DeepSeek V4 Pro + guideline
TACO, TAPE, TAUS	DeepSeek V4 Pro + guideline

The local submission keeps the same assignment and replaces each DeepSeek V4 Pro route with prompted DeepSeek-R1-Qwen-14B.

Spread measured reaction; prediction balance filtered the route choice

Guideline-conditioned system	TACO Argument %	TAPE Argument %	TAUS Argument %	Spread (points)
Qwen2.5-0.5B FT — skewed	37.1	96.2	99.7	62.6
Qwen2.5-Coder-1.5B FT	5.6	29.4	52.9	47.4
SmolLM2-1.7B FT	67.4	40.3	69.1	28.8
TinyLlama-1.1B FT	4.7	8.2	18.8	14.1
DeepSeek-R1-Qwen-14B FT	8.5	7.1	13.8	6.8
Mistral-7B FT	10.0	15.6	15.6	5.6
Llama-3.1-8B FT	12.4	14.4	17.6	5.3
DeepSeek-R1-Qwen-1.5B FT	0.9	0.3	0.3	0.6
DeepSeek V4 Pro prompted — Hybrid	**50.9**	**35.6**	**47.9**	**15.3**
DeepSeek-R1-Qwen-14B prompted — Local	**50.9**	**37.9**	**42.9**	**12.9**

Routing trades many fewer misses for more false alarms on Twitter

Run	True Positive	False Positive	False Negative	True Negative
solo	108	32	292	588
hybrid	323	134	77	486
local	312	136	88	484

Twitter scope: 1,020 evaluations · gold Argument = 400 · gold No-Argument = 620

Exact triplets separate stable cases from genuine label shifts

Gold pattern group	Triplets	solo	hybrid	local
Same-label	221	159 (71.9%)	184 (83.3%)	185 (83.7%)
Shifted	**119**	**0 (0.0%)**	**10 (8.4%)**	**11 (9.2%)**
A/N/N only	**67**	**0 (0.0%)**	**7 (10.4%)**	**9 (13.4%)**

A = Argument · N = No-Argument · order = TACO / TAPE / TAUS

Routing helps overall, but some selected routes fail on test

FINARG

Gold Argument rate
50.0%

Routed prediction rate
60.6%

73 false positives · 37 false negatives

Development route reversals on test

ARGUMINSCI	−8 net correct
PE	−6
USELEC	−5

Routing transfers in aggregate; individual development choices can still fail.

We next tested what the surrounding passage could add

FOUR DATASETS WITH COMPLETE PASSAGES

ABSTRCT · clinical-trial abstracts

PE · persuasive student essays

SCIARK · multidisciplinary scientific abstracts

FINARG · earnings-call Q&A

340 test sentences per dataset

MATCHED INPUTS

1 Target sentence only · no surrounding text

2 Target + original passage · complete source text

3 Target + constrained roles · claim / premise / none

4 Target + open roles · roles named by the generator

Same ModernBERT classifier family

The target sentences, labels, data splits, and seeds stayed fixed.

DeepSeek turned each passage into a source-traceable abstraction/intermediate form

Original passage

Ordered source sentences

Target sentence marked

DeepSeek V4 Flash
→

No gold label
No dataset name
No guideline

GENERATED RECORD

Abstain + reason — explain when the whole passage cannot be processed

Summary + role definitions — overview of the generated schema

Unit text + role — grounded content and its function

Source sentence IDs + quotes — link every unit to the passage

Preserved features — numbers, entities, negation, uncertainty, comparisons, causes

Coverage notes + risk flags — record omissions and review warnings

Generated once per passage; reused as classifier input and as a role-based decision cue.

The guideline helps, but exact rule following remains rare

The two follow-up rows use the same 340 tweets and V4 Pro setup; Hybrid is the official historical reference.

GUIDELINE ONLY (Hybrid)

Tweet + active guideline

0.782

Main macro-F1

10 / 119

shifted triplets fully correct

SUBMITTED HYBRID TWITTER ROUTE

CONTEXT ONLY

Tweet + passage · no guideline

0.743

Main macro-F1

0 / 119

shifted triplets fully correct

POST-RELEASE FOLLOW-UP

CONTEXT + GUIDELINE

Tweet + passage + active guideline

0.776

Main macro-F1

10 / 119

shifted triplets fully correct

POST-RELEASE FOLLOW-UP

Matched graph structure helps only against its own pooling control

System	Equal-dataset mean	vs matched pool
Original-passage pool (OC-POOL)	0.565	—
Original-passage graph (OC-GAT)	0.592	+0.027
Constrained pool (CU-POOL)	0.575	—
Constrained graph (CU-GAT)	**0.718**	**+0.143**
Open pool (OU-POOL)	0.588	—
Open graph (OU-GAT)	0.624	+0.036

Only the matched 0.575 → 0.718 pooling comparison isolates the value of recorded links.
This does not mean the graph recovers semantic argument relations.

Four-dataset Touché comparison and post-release construction

Dataset	Submitted Hybrid	Experiment 3 ensemble	Selected input
ABSTRCT	0.900	0.918	Constrained structure
PE	0.771	0.844	Open structure
SCIARK	0.823	0.865	Original passage + guideline
FINARG	0.673	0.718	Constrained structure

POST-RELEASE TEN-DATASET CONSTRUCTION · NOT SUBMITTED

Keep Hybrid predictions on the other six Supplementary datasets.
Replace only ABSTRCT, PE, SCIARK, and FINARG with the development-selected ensembles.
Result: 0.8305 versus submitted Solo 0.8148 and submitted Hybrid 0.813.
Different training and inputs; not an official score and not evidence that generated structure caused the increase.

Original passages were strongest in the grouped flat-input comparison

Equal-dataset cross-dataset macro-F1 · same flat ModernBERT classifier

INPUT	MEAN MACRO-F1
Sentence only	0.562
Constrained generated structure	0.681
Open generated structure	0.684
Original passage	**0.695**

Generated structures retained much of the context gain, but did not beat the original passage.

Roles-only input is stronger than generated text or complete structure

ModernBERT input from Experiment 3 records	Equal-dataset mean
Generated text only (roles removed)	0.667
Complete generated structure	0.681
Generated roles only	**0.741**
Original passage, for reference	0.695

This is the DeepSeek Experiment 3 diagnostic, not a Twitter result.
Qwen's matched role-only student is 0.755 and agrees with its fixed role rule.
Internal codes: CU-T0 0.667, CU-F0 0.681, CU-RO 0.741, OC 0.695.
A generated role remains a model output, not human gold.