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

Touché at CLEF 2026

Muhammad Qasim Hussain[1,*], Maximilian Heinrich[1], Midhun Kanadan[1] and Benno Stein[1]

[1]Bauhaus-Universität Weimar, Weimar, Germany

Abstract

Argument identification — deciding whether a sentence expresses an argument — is commonly treated as sentence-level binary classification. This framing hides a central difficulty: whether a sentence counts as an argument depends on the annotation guideline. A model trained on sentence text alone therefore cannot transfer reliably across corpora that define arguments differently. The Touché 2026 shared task on *Generalizability of Argument Identification in Context* makes this difficulty explicit, evaluating ten in-domain corpora alongside three out-of-domain Twitter variants that reuse the same sentences under different guidelines. We compare off-the-shelf and fine-tuned LLMs and examine how access to the annotation guideline changes their predictions. Guideline conditioning helps the strongest mid-sized fine-tuned LLMs but is less reliable for smaller models, and no single system is best everywhere. We therefore introduce dataset-level expert routing, sending each corpus to the system that is strongest on it. On the official test, the strongest single model performs best in-domain but transfers poorly to the out-of-domain Twitter variants. Routing across complementary systems closes much of this gap and yields our strongest out-of-domain result, highlighting both the value and the limits of guideline-conditioned expert selection.

Keywords

Argument Mining, Argument Identification, Cross-Domain Generalization, Expert Routing

1. Introduction

Argumentation is a basic part of how people reason, justify decisions, and disagree with one another. It appears in scientific articles, political debates, financial commentary, and everyday social-media exchanges, and it supplies the structured reasoning that links claims to the evidence offered for them. Making this structure accessible to machines is the goal of *argument mining*, the computational study of argumentative text. Its first and most fundamental step is *argument identification*: deciding whether a span of text expresses an argument at all. Only once argumentative spans have been located can later stages classify their roles or the relations between them [1, 2].

Argument identification is commonly framed as sentence-level binary classification: given a sentence, decide whether it is argumentative. This framing is convenient, but it conceals a central difficulty in cross-domain settings. A sentence is not argumentative purely by virtue of its wording; it is argumentative *relative to an annotation guideline*. When corpora differ not only in genre but also in what counts as an argument, a sentence-only model is forced to learn a single averaged decision boundary across incompatible criteria. The Touché 2026 shared task on *Generalizability of Argument Identification in Context* makes this difficulty explicit [3, 4]. Beyond ten in-domain corpora, it evaluates systems on out-of-domain Twitter variants that reuse the same underlying sentences under different annotation guidelines. A classifier that observes only the sentence cannot distinguish these cases and must assign

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

muhammad.qasim.hussain@uni-weimar.de (M. Q. Hussain); maximilian.heinrich@uni-weimar.de (M. Heinrich); midhun.kanadan@uni-weimar.de (M. Kanadan); benno.stein@uni-weimar.de (B. Stein)

0009-0009-2007-6452 (M. Q. Hussain); 0000-0001-5450-8203 (M. Heinrich); 0009-0008-7030-9878 (M. Kanadan); 0000-0001-9033-2217 (B. Stein)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

repeated sentences the same label regardless of the active guideline.

We study two questions: whether providing the annotation guideline improves argument identification over sentence-only input, and whether it improves out-of-domain behavior when the same sentences are judged under different criteria. We evaluate two LLM setups, *fine-tuned LLMs* and *off-the-shelf LLMs*, each with sentence-only input and with the guideline supplied as an explicit instruction.

These experiments yield two findings that shape our submission: guideline conditioning helps the stronger fine-tuned LLMs, but no single system is best on every corpus. We therefore route each dataset to the system with the strongest development evidence. This dataset-level expert routing is our main practical contribution, and it yields our best out-of-domain result.

The remainder of this paper is organized as follows. Section 2 reviews related work on argument mining and cross-domain generalization. Section 3 introduces the task, data, and the three system types we compare. Section 4 reports the development results, showing which systems are strongest in-domain and which actually react to a changed guideline out-of-domain. Section 5 turns these results into the dataset-level routing decision and reports the official test results. Section 6 analyzes the remaining errors after the release of the test labels. Section 7 summarizes the main findings.

2. Related Work

Argument mining seeks to detect argumentative units and their relations in text. Surveys by Cabrio and Villata [1] and Lawrence and Reed [2] trace how rapidly the field has diversified across tasks, domains, and evaluation settings. The former organize the task as a pipeline in which argument extraction precedes relation prediction, so that identifying argumentative spans is the entry point on which later stages depend. The latter caution that this pipeline view oversimplifies the interdependencies between subtasks and advocate a more networked perspective; even so, they retain an argument/non-argument decision as an early step. Either way, deciding whether a span is argumentative comes first, and errors made there propagate through everything that follows, which motivates our focus on identification in isolation.

Argument mining has been applied to a wide range of datasets that span very different genres and domains. Persuasive essays offer clean, explicitly structured argumentation [5], while debate and policy material contributes a more spoken, adversarial, and institutionally framed style [6]; large integrated resources such as IAM combine several argument-mining subtasks at scale [7]. Scientific corpora—from computer-graphics publications [8] to biomedical abstracts and clinical-trial reports [9, 10]—tie argumentative signals to evidence, study design, and disciplinary writing conventions. Comparative web sentences, online dialogue, and financial earnings calls add yet another mix: implicit comparisons, conversational fragments, and domain-specific claims that depart from textbook argumentative prose [11, 12, 13, 14]. Across this collection, "argument" is operationalized differently from one annotation project to the next—a divergence that proves consequential under cross-domain evaluation.

Daxenberger et al. [15] provide early empirical evidence for this difficulty. Across six claim-identification datasets, they find that annotation schemes conceptualize claims differently, which harms cross-domain transfer even when some lexical properties and system configurations transfer. More recently, Feger et al. [16] show that strong models tend to learn dataset identity and dataset-specific regularities rather than a portable notion of argumentativeness. The Touché 2026 task builds on this observation, evaluating systems across ten in-domain corpora and three out-of-domain Twitter variants [3, 4]. The Twitter setting is especially diagnostic: TACO supplies naturally occurring conversational tweets [17], and the shared task relabels the same sentences under different guideline definitions. When the input sentence is held fixed and only the guideline changes, any correct label change must come from modeling the annotation criterion rather than the text alone—an ability that sentence-only classifiers structurally lack.

Methodologically, systems for this task have so far rested largely on pretrained transformers, so it is worth situating our approach against the model families it builds on. Transformer encoders remain a strong baseline for sentence classification. The pretrain-then-fine-tune recipe of BERT-style

models [18, 19] underpins much supervised text classification, including argument identification, where a fully fine-tuned, sentence-only encoder maps each sentence to a label independently of any annotation guideline. Indeed, such encoders remain the strongest reported systems on individual argument-identification benchmarks [16]. This design is effective within a single dataset, but it fixes one label per sentence: the guideline is baked into the learned weights rather than exposed as an input.

Instruction-tuned language models offer a different route: the annotation guideline can be supplied as input rather than collapsed into a fixed label mapping, which makes guideline-conditioned classification practical. A growing set of open-weight models across a wide range of scales makes this route feasible even under limited compute; we describe the specific models, parameter-efficient adaptation, and serving we use in Section 3.

3. Experimental Setup

This section describes our experimental setup in three parts. We first introduce the shared task and its data, then define the system types we compare, and finally detail the training protocol behind each system.

3.1. Touché Dataset and Task

The Touché 2026 shared task on *Generalizability of Argument Identification in Context* [3, 4] is a binary sentence-classification problem. Each instance provides a sentence together with provenance metadata such as the source document and the dataset-specific annotation guideline, and the system must predict one of the two labels `Argument` or `No-Argument`.

It draws on ten established public benchmark datasets for argument identification: biomedical trial abstracts (ABSTRCT) [10], comparative web sentences (ACQUA) [11], online-forum dialogue posts (AEC, AFS) [12, 13], scientific publications and abstracts (ARGUMINSCI, SCIARK) [8, 9], financial earnings-call transcripts (FINARG) [14], Wikipedia debating topics (IAM) [7], persuasive essays (PE) [5], and US presidential-debate transcripts (USELEC) [6]. In the shared-task release, each in-domain corpus contributes 1,020 training, 340 development, and 340 official test instances. The train and development splits are exactly balanced, with 510 `Argument` and 510 `No-Argument` instances in training and 170 instances of each class in development. Across the ten in-domain corpora, this yields 10,200 training instances and 3,400 development instances overall. Official test labels are not available during system development.

Out-of-domain evaluation includes three Twitter variants derived from TACO [17]: TACO, TAPE, and TAUS. These variants reuse the same underlying Twitter sentences but pair them with three different label definitions: TACO keeps the original TACO labels, TAPE applies labels aligned with the PE guideline plus Twitter-specific deviations, and TAUS applies labels aligned with the USELEC guideline plus Twitter-specific deviations. The key difficulty is that the same sentence can receive different labels under different guidelines.

3.2. Experimental Design

Our experiments proceed in two stages. We first compare off-the-shelf and fine-tuned LLM setups on the ten in-domain benchmark corpora, contrasting sentence-only with guideline-conditioned input where applicable. We then analyze out-of-domain behavior on the three Twitter variants introduced above, focusing on whether guideline-conditioned prompting changes predictions under shifted annotation guidelines.

Across these experiments, we distinguish two LLM setups. *Off-the-shelf LLMs* are instruction-following language models that we do not fine-tune; we query them through prompts only, either locally or through a remote API. *Fine-tuned LLMs* are generative large language models that we fine-tune on the task with parameter-efficient QLoRA. For comparison, we also include a fully fine-tuned RoBERTa

sentence-only baseline as a reference to the encoder-classifier line of prior argument-identification work.

3.3. Training Protocol

We train all fine-tuned systems under a shared protocol and describe the data, each model group, and the encoder baseline in turn below.

Training data. For the ten in-domain corpora we use the shared-task splits unchanged: fine-tuned systems are trained on the official train split and scored on the held-out development split. No TACO, TAPE, or TAUS Twitter sentences are used for fine-tuning.

Fine-tuned LLMs. This group comprises three 7B–14B models as our main systems — `Mistral-7B` [20], `Llama-3.1-8B` [21], and `DeepSeek-R1-14B` [22] — and, to probe the effect of model scale on guideline conditioning, five additional sub-2B LLMs: `TinyLlama-1.1B` [23], `SmolLM2-1.7B` [24], `Qwen2.5-Coder-1.5B` [25], `DeepSeek-R1-1.5B` [22], and `Qwen2.5-0.5B` [26]. Throughout, `DeepSeek-R1-14B` and `DeepSeek-R1-1.5B` abbreviate the `DeepSeek-R1-Distill-Qwen-14B` and `DeepSeek-R1-Distill-Qwen-1.5B` checkpoints. For the fine-tuned LLMs, we train separate QLoRA runs for sentence-only and guideline-conditioned input. In the sentence-only setup, the model receives only the target sentence; in the guideline-conditioned setup, it receives the dataset guideline together with the target sentence. Both runs use the same labels and the same train/dev split, but produce separate adapters and checkpoints. All models in this setup are adapter-fine-tuned with QLoRA, using 4-bit NF4 quantization [27], low-rank adapters [28], three training epochs, and a target effective batch size of 512 examples per optimizer step. `TinyLlama-1.1B` receives a compact guideline variant in its guideline-conditioned runs because its 2048-token context cannot accommodate the full guideline prompt. Learning rates and adapter targets are model-specific.[1]

Off-the-shelf LLMs. This category comprises `DeepSeek-V4-Pro` [29], accessed via the DeepSeek API, and `DeepSeek-R1-14B` [22], served locally with vLLM [30]. We evaluate both off-the-shelf models with sentence-only and guideline-conditioned prompts. The local `DeepSeek-R1-14B` run uses the same OpenAI-compatible chat-inference setup as the API model, but without task-specific fine-tuning.

RoBERTa baseline. As an encoder reference point against the LLM setups, we also train a fully fine-tuned RoBERTa sentence encoder [19] for three epochs on sentence-only input, with a smaller effective batch size and a shorter maximum sequence length than in the QLoRA runs. RoBERTa enters our comparison only in the per-corpus development scores of Table 2 and does not participate in the out-of-domain diagnostic or the routing decision.

4. Results

Our development results characterize our systems in two complementary ways. The in-domain corpora identify the strongest system for each corpus, while the out-of-domain Twitter variants reveal which systems actually react to a changed annotation guideline.

4.1. In-Domain Results

Table 1 contrasts sentence-only with guideline-conditioned input where paired data are available, isolating the effect of input format on pooled macro-F_1. Table 2 then breaks down the strongest fine-tuned LLM runs and RoBERTa by corpus to expose dataset-specific strengths.

[1] The Qwen-based runs use a more conservative setup than the Mistral and Llama runs, namely a lower learning rate, FP16 computation, and attention-only adapters; the exact settings are listed in Appendix A.

Table 1
Paired sentence-only vs. guideline-conditioned comparison, evaluated on the shared task's development set and grouped by system type. *Sentence* (sentence-only) gives the system only the target sentence; *Guideline* (guideline-conditioned) additionally supplies the dataset guideline — for fine-tuned LLMs as two separate fine-tuning runs and for the off-the-shelf LLMs as two prompt formats. *Change* is *Guideline* minus *Sentence*, in percentage points (pp). Bold marks the best value in each column. Within each system-type group, systems are ordered by descending *Guideline*. `TinyLlama-1.1B` uses a compact guideline variant; RoBERTa is sentence-only in this study and therefore has no paired guideline-conditioned entry. `DeepSeek-R1-14B` and `DeepSeek-R1-1.5B` abbreviate the `DeepSeek-R1-Distill-Qwen-14B` and `DeepSeek-R1-Distill-Qwen-1.5B` checkpoints.

System	Sentence	Guideline	Change (pp)
Fine-tuned LLMs			
`Mistral-7B`	**0.760**	**0.815**	+5.5
`Llama-3.1-8B`	0.760	0.800	+4.0
`DeepSeek-R1-14B`	0.703	0.736	+3.3
`TinyLlama-1.1B` (compact)	0.730	0.684	−4.6
`SmolLM2-1.7B`	0.656	0.636	−2.0
`Qwen2.5-Coder-1.5B`	0.561	0.625	+6.4
`DeepSeek-R1-1.5B`	0.496	0.603	**+10.7**
`Qwen2.5-0.5B`	0.546	0.595	+4.9
Off-the-shelf LLMs			
`DeepSeek-V4-Pro`	0.659	0.685	+2.6
`DeepSeek-R1-14B`	0.655	0.682	+2.7

Table 2
Per-corpus development macro-F_1, grouped by system type. The table includes the RoBERTa baseline, the strongest fine-tuned LLM runs from Table 1, and two guideline-prompted off-the-shelf reference models. *Guid* and *Sent* mark the guideline-conditioned and sentence-only input format of the run. The off-the-shelf rows are guideline-prompted rather than fine-tuned. The corpus abbreviations are ABS = ABSTRCT, ACQ = ACQUA, AEC = AEC, AFS = AFS, ARG = ARGUMINSCI, FIN = FINARG, IAM = IAM, PE = PE, SCI = SCIARK, and USE = USELEC. *Pool* denotes pooled macro-F_1 over the full development set. Bold marks the best value in each column. Within each system-type group, systems are ordered by descending *Pool*. `DeepSeek-R1-14B` abbreviates the `DeepSeek-R1-Distill-Qwen-14B` checkpoint.

System	ABS	ACQ	AEC	AFS	ARG	FIN	IAM	PE	SCI	USE	Pool
RoBERTa	0.891	0.774	0.671	0.670	0.808	0.723	0.726	0.711	0.769	0.703	0.746
Fine-tuned LLMs											
`Mistral-7B` Guid	**0.923**	0.822	**0.965**	**0.838**	0.832	**0.728**	**0.779**	0.750	**0.812**	0.700	**0.815**
`Llama-3.1-8B` Guid	0.921	0.796	0.894	0.820	0.835	0.715	0.744	0.731	0.809	**0.732**	0.800
`Llama-3.1-8B` Sent	0.921	0.720	0.682	0.744	**0.847**	0.696	0.694	**0.779**	**0.812**	0.705	0.760
`DeepSeek-R1-14B` Guid	0.858	0.786	0.599	0.762	0.762	0.675	0.715	0.712	0.791	0.691	0.736
Off-the-shelf LLMs											
`DeepSeek-V4-Pro` Guid	0.718	**0.846**	0.562	0.769	0.724	0.650	0.694	0.579	0.640	0.594	0.685
`DeepSeek-R1-14B` Guid	0.711	0.829	0.567	0.692	0.667	0.665	0.728	0.652	0.585	0.671	0.682

The fully fine-tuned RoBERTa baseline does not match the strongest guideline-conditioned fine-tuned LLMs, but remains competitive under sentence-only input: it reaches 0.746 pooled macro-F_1, within 0.014 of sentence-only `Mistral-7B` and ahead of most sub-2B LLMs. The paired comparison in Table 1 shows the clearest gains for the larger fine-tuned models we tested: guideline conditioning improves `Mistral-7B` from 0.760 to 0.815, `Llama-3.1-8B` from 0.760 to 0.800, and `DeepSeek-R1-14B` from

Table 3
Predicted `Argument` percentage of each system on the three out-of-domain Twitter variants, which reuse the same sentences under different annotation guidelines. *Spread* is the maximum minus the minimum predicted-`Argument` rate across TACO, TAPE, and TAUS, in percentage points (pp); a spread of 0 means the system predicts identically regardless of the active guideline. The table compares guideline-conditioned LLM runs only. *Guid* denotes guideline-plus-sentence input — realized as QLoRA fine-tuning for the fine-tuned rows and as prompting for the off-the-shelf rows — and `TinyLlama-1.1B` uses a compact guideline variant. Within each system-type group, systems are ordered by descending *Spread*. `DeepSeek-R1-14B` and `DeepSeek-R1-1.5B` abbreviate the `DeepSeek-R1-Distill-Qwen-14B` and `DeepSeek-R1-Distill-Qwen-1.5B` checkpoints.

System	TACO (%)	TAPE (%)	TAUS (%)	Spread (pp)
Fine-tuned LLMs				
`Qwen2.5-0.5B` Guid	37.1	96.2	99.7	62.6
`Qwen2.5-Coder-1.5B` Guid	5.6	29.4	52.9	47.4
`SmolLM2-1.7B` Guid	67.4	40.3	69.1	28.8
`TinyLlama-1.1B` Guid	4.7	8.2	18.8	14.1
`DeepSeek-R1-14B` Guid	8.5	7.1	13.8	6.8
`Mistral-7B` Guid	10.0	15.6	15.6	5.6
`Llama-3.1-8B` Guid	12.4	14.4	17.6	5.3
`DeepSeek-R1-1.5B` Guid	0.9	0.3	0.3	0.6
Off-the-shelf LLMs				
`DeepSeek-V4-Pro` Guid	50.9	35.6	47.9	15.3
`DeepSeek-R1-14B` Guid	50.9	37.9	42.9	12.9

0.703 to 0.736. Among all single systems, guideline-conditioned `Mistral-7B` is best at 0.815 pooled macro-F_1.

The effect is not universal, however, and smaller models often benefit less or become less stable when the guideline is added. `TinyLlama-1.1B` drops from 0.730 to 0.684 and `SmolLM2-1.7B` from 0.656 to 0.636, whereas `Qwen2.5-Coder-1.5B` and `DeepSeek-R1-1.5B` gain more strongly from guideline conditioning, albeit from much lower baselines. Without task-specific fine-tuning, the off-the-shelf guideline-prompted references reach 0.685 pooled macro-F_1 for `DeepSeek-V4-Pro` Guid and 0.682 for `DeepSeek-R1-14B` Guid — below all three fine-tuned 7B–14B LLMs and RoBERTa — and match or exceed the fine-tuned models only on individual corpora.

Looking per corpus, the development scores in Table 2 show that the strongest overall model is not best everywhere. Guideline-conditioned `Mistral-7B` leads or ties on most in-domain corpora, but sentence-only `Llama-3.1-8B` wins ARGUMINSCI and PE, guideline-conditioned `Llama-3.1-8B` wins USELEC, and `DeepSeek-V4-Pro` Guid wins ACQUA.

4.2. Out-of-Domain Twitter Diagnostics

The out-of-domain evaluation uses the three Twitter variants TACO, TAPE, and TAUS introduced in Section 3.1. Their development labels are hidden, so we cannot compute macro-F_1. Instead, we use the fact that the variants share identical sentences and differ only in their annotation guideline: any change in predicted `Argument` rate must come from the system reacting to the active guideline.

We summarize this reaction in a single number, the *spread* — the maximum minus the minimum predicted-`Argument` rate across TACO, TAPE, and TAUS. A spread near 0 means the system predicts identically regardless of the guideline; a larger spread indicates that predictions change when the guideline changes. The spread tells us whether predictions move with the guideline, but not whether they move in the right direction, so we treat it as informative only when the resulting class balance remains plausible. Sentence-only systems are omitted from Table 3 for exactly this reason: by construction they would have spread 0.

Under this diagnostic, the off-the-shelf LLMs `DeepSeek-V4-Pro` Guid and `DeepSeek-R1-14B` Guid

Table 4
Dataset-level expert routing behind submission `hybrid`, our cloud-assisted run. *Guid* and *Sent* denote guideline-conditioned and sentence-only input. `Mistral-7B` and `Llama-3.1-8B` are the fine-tuned experts, whereas `DeepSeek-V4-Pro` is off-the-shelf. Rows follow the model order of Table 2 (descending pooled macro-F_1 of the routed expert), with the ten in-domain corpora listed first and the three out-of-domain Twitter variants at the end. Submission `local` reuses the same assignment with every `DeepSeek-V4-Pro` Guid route replaced by `DeepSeek-R1-14B` Guid; both are off-the-shelf LLMs, used only via prompting. `DeepSeek-R1-14B` abbreviates the `DeepSeek-R1-Distill-Qwen-14B` checkpoint.

Dataset(s)	Routed expert
ABSTRCT, AEC, AFS, FINARG, IAM, SCIARK	`Mistral-7B` Guid
USELEC	`Llama-3.1-8B` Guid
ARGUMINSCI, PE	`Llama-3.1-8B` Sent
ACQUA	`DeepSeek-V4-Pro` Guid
TACO, TAPE, TAUS	`DeepSeek-V4-Pro` Guid

Table 5
Official Touché 2026 test results for our three submitted runs. All values are macro-F_1. TACO, TAPE, and TAUS are the three out-of-domain Twitter variants; *Main* is their unweighted mean and is the official ranking metric, while *Supplementary* is the unweighted mean of macro-F_1 over the ten in-domain corpora. `solo` is a single guideline-conditioned `Mistral-7B` model, `hybrid` is the cloud-assisted dataset-level expert routing, and `local` is the API-free routing variant. Runs are listed in descending order of the *Main* score. Bold marks the best value in each column.

Run	TACO	TAPE	TAUS	Main	Supplementary
`hybrid`	**0.827**	0.747	**0.774**	**0.782**	0.813
`local`	0.791	**0.751**	0.766	0.769	0.806
`solo`	0.510	0.622	0.650	0.594	**0.815**

produce the largest spreads with plausible class balance (15.3 and 12.9 percentage points), whereas the fine-tuned 7B–14B guideline LLMs show much smaller shifts — about 5–7 percentage points — and remain close to their original class balance. Several small fine-tuned LLMs show much larger spreads, but these are often accompanied by highly skewed predicted class balances; for example, `Qwen2.5-0.5B` predicts `Argument` for 96.2% of TAPE and 99.7% of TAUS.

5. Submissions and Official Results

The results in Section 4 motivate a routed submission design: different datasets favor different systems, so the submission follows the strongest development evidence for each case. We submit three complementary runs through TIRA [31], the official submission platform of the shared task.

The first run, `solo`, uses our strongest single model, the guideline-conditioned `Mistral-7B`. The second, `hybrid`, routes each dataset to its strongest development system: per-corpus macro-F_1 for the ten in-domain corpora (Table 2) and the spread diagnostic for the three Twitter variants (Table 3). The third run, `local`, follows the same routing but replaces `DeepSeek-V4-Pro` Guid with `DeepSeek-R1-14B` Guid, so that the full run can execute without an API. Table 4 summarizes the resulting assignments.

5.1. Official Test Results

Table 5 reports the official test scores of our three runs. The shared task uses two complementary score aggregations. The *Main* score, which determines the ranking, is the unweighted mean of macro-F_1 over the three out-of-domain Twitter variants. The *Supplementary* score is the corresponding unweighted mean over the ten in-domain corpora. The official scores use different aggregations from the pooled development macro-F_1 in Section 4.

Table 6
Post-release error analysis on the three Twitter variants. The table uses the released test labels and reports pooled Twitter behavior across TACO, TAPE, and TAUS. *Pred. Arg.* is the predicted `Argument` rate, *Prec.* and *Rec.* are precision and recall for the `Argument` class. Because scores are pooled over all three variants, the *Macro-*F_1 column is computed differently from the per-variant mean reported as the Main score in Table 5 and therefore differs slightly from it.

Run	Pred. Arg.	Macro-F_1	Prec.	Rec.
`solo`	0.137	0.592	0.771	0.270
`hybrid`	0.448	0.788	0.707	0.807
`local`	0.439	0.774	0.696	0.780

Our best official result comes from `hybrid`, which reaches a Main score of 0.782. `local` follows closely at 0.769, showing that running the same routing entirely on local models lowers the official Main score by only 0.013; this small gap is consistent with the development diagnostics, where `DeepSeek-R1-14B` Guid is slightly weaker than `DeepSeek-V4-Pro` Guid in-domain but shows similar guideline sensitivity on the Twitter variants. The `solo` run, by contrast, is strongest in-domain, reaching a Supplementary score of 0.815, but has the lowest Main score among our runs, 0.594. This mirrors the development pattern: the strongest single model remains in-domain while dropping sharply on the out-of-domain variants, where its per-corpus macro-F_1 falls to 0.510–0.650. Most of this out-of-domain gap is recovered by the routed runs. On the full leaderboard, `hybrid` ranks fourth on the official Main metric, 0.013 behind the leading run.

The remaining differences help explain the gap between the runs. On ARGUMINSCI, PE, and USELEC, the development-based decision to route away from the guideline-conditioned `Mistral-7B` did not pay off in the official test results, where the single model would have scored slightly higher. This mismatch shows that not every per-corpus routing advantage is equally stable. The ACQUA route, in turn, explains most of the gap between the two hybrids: `hybrid` uses `DeepSeek-V4-Pro` Guid for this dataset and reaches 0.862, whereas `local` replaces it with `DeepSeek-R1-14B` Guid and reaches 0.791.

6. Error Analysis

The released test labels allow us to separate two effects that were only partly visible on development data: whether routing improves the Twitter score, and whether it actually tracks guideline-dependent label changes. The answer is mixed. Routing substantially improves the official Main score, but mostly because it corrects the single fine-tuned model's low recall on Twitter. It does not yet solve the harder problem of deciding when the same sentence should change label under a different guideline.

The two official scores of the `solo` run make this trade-off concrete. The same conservative decision boundary that yields its strong in-domain Supplementary score (0.815) depresses its Main score to 0.594: balanced in-domain corpora reward precise `Argument` predictions, whereas the positive-heavy out-of-domain Twitter variants penalize the model's reluctance to predict them. In-domain strength and out-of-domain weakness are thus two sides of one model property, which is why the routed runs recover the Main score without changing the in-domain experts.

Table 6 shows the class-balance problem. Across TACO, TAPE, and TAUS, 39.2% of the released test instances are labeled `Argument`. The `solo` run predicts that label for only 13.7% of instances. Its positive predictions are therefore often correct, but it misses most Twitter arguments. The routed runs replace this conservative Twitter behavior with guideline-prompted off-the-shelf models and move much closer to the gold class balance. As a result, `Argument` recall rises from 0.270 for `solo` to 0.807 for `hybrid` and 0.780 for `local`. The Main-score gain is therefore primarily a recall gain rather than a broad reduction of all error types.

This recall gain does not imply that the systems model guideline shifts correctly. To test that stricter claim, we group repeated Twitter sentences into triplets, one prediction each for TACO, TAPE, and TAUS,

Table 7
Exact recovery of guideline-dependent label patterns after test labels were released. Each Twitter sentence recurs in all three out-of-domain variants, so its three gold labels form a triplet, written in the order TACO/TAPE/TAUS with A = Argument and N = No-Argument (for example, A/N/N); eight such patterns are possible. *Same-label triplets* are the two patterns whose label is constant across guidelines, A/A/A and N/N/N; *shifted triplets* are the remaining six patterns, in which the gold label changes with the guideline. The last row isolates A/N/N, the largest shifted group. The *Triplets* column gives the number of gold triplets in each group; each run's entry is the number of triplets it recovers exactly — all three predictions matching the gold pattern — with the within-group rate in parentheses.

Gold pattern group	Triplets	`solo`	`hybrid`	`local`
Same-label triplets	221	159 (0.719)	184 (0.833)	185 (0.837)
Shifted triplets	119	0 (0.000)	10 (0.084)	11 (0.092)
A/N/N only	67	0 (0.000)	7 (0.104)	9 (0.134)

written in that order with A for Argument and N for No-Argument. Table 7 reports exact pattern recovery for these triplets. When all three guidelines agree, the routed systems recover 184–185 of 221 same-label triplets (gold pattern A/A/A or N/N/N). The shifted triplets are much harder: among the 119 cases where at least one guideline assigns a different label, hybrid recovers only 10 triplets exactly and local recovers 11. The largest shifted group is A/N/N, where a sentence is argumentative under the native Twitter guideline TACO but not under TAPE or TAUS. The solo model usually misses these cases as N/N/N; the routed systems often overcorrect and assign too many Argument labels. Routing therefore fixes the low-recall failure, but it can become over-positive on stricter Twitter variants, especially when a tweet expresses topical stance without satisfying the active guideline.

Appendix C collects qualitative examples that illustrate this trade-off. The routed system recovers straightforward positives and some true guideline shifts — for example, a #Brexit tweet missed by solo but labeled A/A/A by the routed runs. Its main failure mode is overextension: topical or stance-like tweets can be treated as arguments even when the active guideline is stricter. Short reactions show the same tendency in an even shorter form: a phrase such as "No way" is non-argumentative under all three guidelines, but is partly read as an argument (Table 9).

The in-domain errors expose a separate limitation: development-based routing is not equally stable for every corpus. On the official test labels, routing away from Mistral-7B hurts slightly on ARGUMINSCI, PE, and USELEC, although the ACQUA route remains beneficial. Among the in-domain corpora, FINARG shows the clearest calibration problem: the routed model predicts Argument for 60.6% of cases while the gold rate is 50.0%, producing 73 false positives and 37 false negatives. Financial statements are therefore often over-read as argumentative.

Overall, the error analysis narrows the interpretation of our submissions. Guideline-aware routing is useful because different datasets favor different systems, but the gains come from a specific correction rather than from fully guideline-sensitive reasoning. For in-domain corpora, small development-set advantages do not always transfer to the blind test. For the Twitter variants, the main unresolved problem is not detecting argumentative language in general; it is deciding whether a stance-like or reply-like sentence satisfies the active guideline. A stronger routed system would therefore need more conservative in-domain expert selection and better calibration for Twitter cases where topical stance alone is not enough for an Argument label.

7. Conclusion

We examined when explicit guideline conditioning improves argument identification in the Touché 2026 shared task on *Generalizability of Argument Identification in Context*. Returning to the two questions from the introduction, our findings on the guideline's effect are mixed. First, supplying the annotation guideline improves argument identification over sentence-only input, but unevenly: the gain is reliable for the stronger mid-sized fine-tuned LLMs and weak or unstable for smaller ones. Second, the guideline

matters most in the out-of-domain Twitter setting, where the same sentences are judged under different criteria; there, strong in-domain effectiveness alone does not transfer, and combining complementary systems through dataset-level routing closes most of the resulting gap.

The post-release error analysis qualifies this result. Routing improves the Twitter score mainly by correcting the single model's positive-recall failure, but exact guideline-dependent relabeling remains difficult when the same sentence changes label across the three Twitter guidelines. Thus, the submissions show that guideline-conditioned routing is useful, but they do not show that current systems fully model annotation guidelines.

A remaining question is how this pattern changes with model scale. In our experiments, guideline conditioning was most consistently beneficial for the stronger mid-sized fine-tuned LLMs, while smaller models behaved more unevenly. Whether larger models can benefit more from stronger prompting or different adaptation strategies deserves further study.

Acknowledgments

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Opus (Anthropic) and GPT-5 (OpenAI) in order to: Drafting content, Paraphrase and reword, Improve writing style, Abstract drafting, Grammar and spelling check, Formatting assistance, Content enhancement. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] E. Cabrio, S. Villata, Five years of argument mining: A data-driven analysis, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, 2018, pp. 5427–5433. URL: https://doi.org/10.24963/ijcai.2018/766. doi:10.24963/ijcai.2018/766.

[2] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2019) 765–818. URL: https://doi.org/10.1162/coli_a_00364. doi:10.1162/coli_a_00364.

[3] J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of Touché 2026 — argumentation systems, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 17th International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[4] J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of Touché 2026 — argumentation systems — extended version, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[5] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Computational Linguistics 43 (2017) 619–659. URL: https://doi.org/10.1162/coli_a_00295. doi:10.1162/coli_a_00295.

[6] S. Haddadan, E. Cabrio, S. Villata, Yes, we can! mining arguments in 50 years of US presidential campaign debates, in: Proceedings of the 57th Annual Meeting of the Association

for Computational Linguistics, 2019, pp. 4684–4690. URL: https://doi.org/10.18653/v1/P19-1463. doi:10.18653/v1/P19-1463.

[7] L. Cheng, L. Bing, R. He, Q. Yu, Y. Zhang, L. Si, IAM: A comprehensive and large-scale dataset for integrated argument mining tasks, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 2277–2287. URL: https://doi.org/10.18653/v1/2022.acl-long.162. doi:10.18653/v1/2022.acl-long.162.

[8] A. Lauscher, G. Glavas, S. P. Ponzetto, An argument-annotated corpus of scientific publications, in: Proceedings of the 5th Workshop on Argument Mining, 2018, pp. 40–46. URL: https://doi.org/10.18653/v1/W18-5206. doi:10.18653/v1/W18-5206.

[9] A. Fergadis, D. Pappas, A. Karamolegkou, H. Papageorgiou, Argumentation mining in scientific literature for sustainable development, in: Proceedings of the 8th Workshop on Argument Mining, 2021, pp. 100–111. URL: https://doi.org/10.18653/v1/2021.argmining-1.10. doi:10.18653/v1/2021.argmining-1.10.

[10] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: Proceedings of the 24th European Conference on Artificial Intelligence, ECAI 2020, 2020, pp. 2108–2115. URL: https://doi.org/10.3233/FAIA200334. doi:10.3233/FAIA200334.

[11] A. Panchenko, A. Bondarenko, M. Franzek, M. Hagen, C. Biemann, Categorizing comparative sentences, in: Proceedings of the 6th Workshop on Argument Mining, 2019, pp. 136–145. URL: https://doi.org/10.18653/v1/W19-4516. doi:10.18653/v1/W19-4516.

[12] R. Swanson, B. Ecker, M. Walker, Argument mining: Extracting arguments from online dialogue, in: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2015, pp. 217–226. URL: https://doi.org/10.18653/v1/W15-4631. doi:10.18653/v1/W15-4631.

[13] A. Misra, B. Ecker, M. Walker, Measuring the similarity of sentential arguments in dialogue, in: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 276–287. URL: https://doi.org/10.18653/v1/W16-3636. doi:10.18653/v1/W16-3636.

[14] A. Alhamzeh, R. Fonck, E. Versmée, E. Egyed-Zsigmond, H. Kosch, L. Brunie, It's time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset, in: Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing, 2022, pp. 163–169. URL: https://doi.org/10.18653/v1/2022.finnlp-1.22. doi:10.18653/v1/2022.finnlp-1.22.

[15] J. Daxenberger, S. Eger, I. Habernal, C. Stab, I. Gurevych, What is the essence of a claim? cross-domain claim identification, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Association for Computational Linguistics, 2017, pp. 2055–2066. URL: https://doi.org/10.18653/v1/D17-1218. doi:10.18653/v1/D17-1218.

[16] M. Feger, K. Boland, S. Dietze, Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Association for Computational Linguistics, 2025, pp. 23900–23915. URL: https://doi.org/10.18653/v1/2025.acl-long.1164. doi:10.18653/v1/2025.acl-long.1164.

[17] M. Feger, S. Dietze, TACO – Twitter arguments from COnversations, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 2024, pp. 15522–15529. URL: https://aclanthology.org/2024.lrec-main.1349/.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/N19-1423. doi:10.18653/v1/N19-1423.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv:1907.11692, 2019. URL: https://arxiv.org/abs/1907.11692.

[20] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril,

T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, arXiv:2310.06825, 2023. URL: https://arxiv.org/abs/2310.06825.

[21] A. Grattafiori, et al., The Llama 3 herd of models, arXiv:2407.21783, 2024. URL: https://arxiv.org/abs/2407.21783.

[22] DeepSeek-AI, DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, arXiv:2501.12948, 2025. URL: https://arxiv.org/abs/2501.12948.

[23] P. Zhang, G. Zeng, T. Wang, W. Lu, TinyLlama: An open-source small language model, arXiv:2401.02385, 2024. URL: https://arxiv.org/abs/2401.02385.

[24] L. B. Allal, A. Lozhkov, E. Bakouch, G. M. Blázquez, G. Penedo, L. Tunstall, A. Marafioti, H. Kydlícek, A. P. Lajarín, V. Srivastav, J. Lochner, C. Fahlgren, X. Nguyen, C. Fourrier, B. Burtenshaw, H. Larcher, H. Zhao, C. Zakka, M. Morlon, C. Raffel, L. von Werra, T. Wolf, SmolLM2: When Smol goes big – data-centric training of a small language model, arXiv:2502.02737, 2025. URL: https://arxiv.org/abs/2502.02737.

[25] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang, A. Yang, R. Men, F. Huang, X. Ren, X. Ren, J. Zhou, J. Lin, Qwen2.5-Coder technical report, arXiv:2409.12186, 2024. URL: https://arxiv.org/abs/2409.12186.

[26] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, arXiv:2412.15115, 2024. URL: https://arxiv.org/abs/2412.15115.

[27] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems 36, 2023. URL: https://arxiv.org/abs/2305.14314.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://arxiv.org/abs/2106.09685.

[29] DeepSeek-AI, DeepSeek-V4: Towards highly efficient million-token context intelligence, arXiv:2606.19348, 2026. URL: https://arxiv.org/abs/2606.19348.

[30] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with PagedAttention, in: Proceedings of the 29th Symposium on Operating Systems Principles, 2023, pp. 611–626. URL: https://doi.org/10.1145/3600006.3613165. doi:10.1145/3600006.3613165.

[31] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous integration for reproducible shared tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://doi.org/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

A. Training Hyperparameters

Table 8 lists the training settings for all fine-tuned systems, grouped by system type. Off-the-shelf models are omitted because they receive no task-specific training.

Table 8
Training hyperparameters by system type. NF4 is the 4-bit NormalFloat quantization format; LoRA is low-rank adaptation and *LoRA rank* its bottleneck dimension; LR is the learning rate; FP16 denotes half-precision (16-bit) floating-point computation; the *effective batch size* is the number of examples per optimizer step after gradient accumulation. Off-the-shelf LLMs have no task-specific training and are therefore omitted. `DeepSeek-R1-14B` and `DeepSeek-R1-1.5B` abbreviate the `DeepSeek-R1-Distill-Qwen-14B` and `DeepSeek-R1-Distill-Qwen-1.5B` checkpoints.

Setting	Value
RoBERTa (full fine-tuning)	
Epochs	3
Learning rate	2×10^{-5}
Maximum length	256 tokens
Label smoothing	0.10
Effective batch size	64
Fine-tuned LLMs (QLoRA, all runs)	
Quantization	4-bit NF4
Epochs	3
Learning-rate schedule	cosine decay, 10% warmup
Weight decay	0.01
LoRA rank	16
LoRA dropout	0.05
Label smoothing	0.10
Effective batch size	512
Input lengths	
Sentence-only runs	Maximum length 256 tokens
Guideline-conditioned runs	Maximum length 4096 tokens
`TinyLlama-1.1B` guideline variant	Compact guideline, maximum length 2048 tokens
Per-model learning rate and adapters	
`Mistral-7B`, `Llama-3.1-8B`	LR 1×10^{-4}, attention + MLP adapters
`DeepSeek-R1-14B`	LR 5×10^{-5}, FP16, attention-only adapters
`TinyLlama-1.1B`, `SmolLM2-1.7B`, `Qwen2.5-Coder-1.5B`	LR 5×10^{-5}, attention-only adapters
`DeepSeek-R1-1.5B`, `Qwen2.5-0.5B`	LR 5×10^{-5}, attention-only adapters

B. Prompt Formats

The experiments use two input formats. In the sentence-only setting, the model receives only the target sentence. In the guideline-conditioned setting, the system message contains the dataset-specific annotation guideline and the user message contains the target sentence.

The following template illustrates the guideline-conditioned prompt; the sentence-only prompt omits the dataset guideline from the system message.

```
SYSTEM:
You are a classifier for the Touche 2026 argument identification task.
Apply the dataset-specific annotation guideline below.

Dataset: <DATASET>
```

```
<DATASET-SPECIFIC GUIDELINE>

Return exactly one label and nothing else:
Argument
or
No-Argument

USER:
Classify the following sentence as Argument or No-Argument, applying the
dataset-specific rubric above.

Sentence: "<TARGET SENTENCE>"
```

C. Qualitative Error Examples

Table 9 provides qualitative examples for the diagnostics discussed in Section 6. The examples are not aggregate evidence; rather, they illustrate the main observed failure modes: recovery of missed Twitter positives, over-positive predictions on shifted Twitter triplets, short-reaction false positives, and consensus misses on in-domain corpora.

Table 9
Additional qualitative error examples from the released test labels, selected to illustrate the main failure modes discussed in Section 6. For Twitter rows, the *Gold* and *Prediction* columns list labels for the same sentence under the three guideline variants in the order TACO/TAPE/TAUS; for example, A/N/N means `Argument` under TACO and `No-Argument` under TAPE and TAUS. The in-domain rows are grouped separately; a dash ("–") in the *Gold* and *Prediction* columns marks their shared labels, which are stated once in the group heading. In the *Prediction* column, `s`, `h`, and `l` abbreviate the `solo`, `hybrid`, and `local` runs. Text is shortened for readability.

Error pattern	Sentence snippet	Gold	Prediction	Interpretation
Twitter variants (labels in TACO/TAPE/TAUS order)				
Missed Twitter argument recovered	"Here is the real consequence of #Brexit … run down some manufacturing …"	A/A/A	s: N/N/N h: A/A/A l: A/A/A	Routed model recovers a positive triplet missed by the single fine-tuned model.
Exact guideline-shift recovery	"@Jeffrey_Johnson And that same money can stop nonsense if it is used wisely"	A/N/N	s: N/N/A h: A/N/N l: A/A/A	API route gets the active guideline contrast exactly; local route over-predicts.
Stance over-labeled across guidelines	"@Jessica_Gross It's OK to not like things …"	A/N/N	s: N/N/N h: A/A/A l: A/A/A	Stance-like wording is extended to stricter guidelines where the gold label is negative.
Wrong variant in shifted triplet	"@Amber_Wright @Mary_Escobar DWP is going to be busy!"	A/N/A	s: N/N/N h: A/N/N l: A/A/N	Both routed systems detect some shift but miss the exact variant-specific pattern.
Short reaction over-labeled	"@Sharon_Cherry @Helen_Jones No way."	N/N/N	s: N/N/N h: A/N/N l: A/A/N	Short disagreement is sometimes read as an argument although no guideline labels it positive.
Rhetorical question over-labeled	"Imagine buhari in #SquidGame will he survive??"	N/N/N	s: N/N/N h: A/N/N l: A/A/A	Question form and topical stance cues can trigger over-positive Twitter predictions.
In-domain misses (gold `Argument`; all three runs predict `No-Argument`)				
Financial explanation missed	"Some of them tend to be accretive to the average gross margin for services …"	–	–	Financial explanations can be treated as descriptive rather than argumentative.
Evidence statement missed	"A 2016 systematic review and meta-analysis found that moderate ethanol consumption does not prolong life …"	–	–	Evidence-like scientific statements can be missed when no explicit recommendation is present.
Debate evidence missed	"The fact is, the take-home pay of a typical American family … is lower than it's been since 1929."	–	–	Debate evidence without an explicit conclusion is often under-predicted.
Policy recommendation missed	"Also needed are a re-envisioning of land-use planning … and a new coalition …"	–	–	Long scientific/policy recommendations remain difficult for all submissions.