Cognitive Perspectives as Supervisory Signals for Argumentative Role Classification in Social Media

Anonymous ACL submission

Abstract

Argumentative discourse on social media encodes not only *what* users claim but *how* they epistemically frame those claims. Transformer-based argument-mining systems tend to rely on surface-level linguistic cues and conversational structure. In doing so, they overlook a deeper layer of signals rooted in epistemic cognition, the ways individuals construct, justify, and communicate knowledge. We introduce a framework that operationalizes epistemic cognition theory as a multi-label perspective taxonomy and demonstrates that these labels function as effective supervisory signals fro argumentative discourse classification. A structured LLM annotation pipeline enriches 1,734 tweets from the TACO argumentative role datasets with perspective annotations. A key design contribution is our conversation-disjoint evaluation protocol, we identify and correct methodological gap in prior conversational argument mining evaluation, where tweet-level cross-validation permits shared conversational context to inflate performance estimates. Under StratifiedGroupKFold evaluation with conversation-level grouping, perspective-enriched embeddings achieve 0.92 macro F1 on four-class argumentative role classification, outperforming text-only baseline by 12.2pp. A controlled comparison against sentiment, stance, and random auxiliary signals under identical conditions establishes signal specificity. All three fall below the unenriched baseline, confirming that the gains are specific to epistemic perspective content rather than generic auxiliary supervision. Cross-dataset evaluation on PHEME rumour detection under Leave-One-Event-Out cross-validation demonstrates framework transferability; perspective enrichment improves macro F1 by 2.4% while reducing cross-event variance by 40.6% indicating stable generalization to unseen news domain.

1 Introduction

In online discourse, a claim is always accompanied by an epistemic framing, whether users justify it with evidence, ground it in personal experience, hedge with uncertainty, or present it as a belief. This distinction between linguistic content and epistemic mode is consequential: two tweets that make superficially similar claims may differ fundamentally in their argumentative function depending on the cognitive perspective from which they are constructed. Yet transformer-based argument mining systems treat argumentative role classification as a purely linguistic phenomenon, relying on surface form and reply structure while ignoring the epistemic cognition underlying argument construction (Habernal and Gurevych, 2017; Lawrence and Reed, 2019).

This framing has two consequences. First, models miss a class of latent signals that are theoretically motivated as relevant to discourse understanding: epistemic cognition research establishes that individuals adopt distinct reasoning modes when constructing arguments (Kuhn and Weinstock, 2002; Hofer and Pintrich, 1997), and these modes leave detectable traces in text. Second, fine-grained argumentative role classification remains disproportionately difficult for minority classes whose surface forms overlap substantially across roles — a gap that lexical and structural features have not closed despite substantial progress in pre-trained language modeling.

We address both limitations by grounding argument mining in epistemic cognition theory. We introduce a six-label perspective taxonomy derived from the psychology of epistemic cognition (Kuhn and Weinstock, 2002; Baron, 2008): **Information Providing**, **Information Seeking**, **Epistemic/Factual**, **Experiential**, **Opinion/Belief**, and **Actively Open-Minded Thinking (AOT)**. These labels capture the *cognitive mode* of a tweet — the

epistemic process through which a user constructs, justifies, and communicates their claim — rather than its content or position.

We instantiate this taxonomy through a structured LLM annotation pipeline that produces multi-label perspective annotations with verbatim evidence spans, confidence scores, and natural language rationales, applied to the TACO dataset (Feger and Dietze, 2024) of argumentative Twitter conversations. A critical design decision distinguishes our framework from prior work: we enforce **conversation-level disjointness** between training and evaluation splits throughout the pipeline. Tweets within the same conversational thread share implicit discourse context — topic framing, register, and rhetorical positioning — that naive tweet-level cross-validation allows to leak across the train-test boundary. We eliminate this by construction using StratifiedGroupKFold with conversation IDs as groups, ensuring that no tweet appears in a fold that contains a conversational neighbor of any training tweet.

Empirically, we demonstrate three findings. Under this rigorous evaluation protocol, perspective-enriched embeddings achieve 0.92 macro F1 on four-class argumentative role classification, outperforming the conversation-disjoint baseline by 12.2pp. A controlled auxiliary signal comparison - testing sentiment, stance, and random binary features under identical architecture and evaluation conditions, establishes that these gains are specific to epistemic perspective signals: sentiment achieves 0.695, stance 0.700, and random features 0.669 macro F1, all below the unenriched baseline of 0.717. Finally, applying the annotation pipeline to PHEME, a rumour detection dataset across five breaking news events, under Leave-One-Event-Out cross-validation, perspective enrichment improves cross-event macro F1 by 2.4pp while reducing inter-event variance by 40%, demonstrating framework transferability to an unseen task and domain.

Our contributions are as follows:

- **Epistemic grounding of argument mining:** We introduce the first integration of epistemic cognition theory into computational argument mining, operationalizing a six-label multi-label taxonomy grounded in the psychology of knowledge and knowing. We demonstrate that perspective distributions are class-discriminative across argumentative roles, with distinct epistemic signatures that explain historical classification difficulty for minority classes.

- **Conversation-disjoint evaluation protocol:** We identify and correct a methodological gap in prior conversational argument mining evaluation: tweet-level cross-validation permits shared conversational context to inflate test performance. We formalize and implement StratifiedGroupKFold evaluation with conversation-level grouping, provide empirical quantification of the performance difference (1.2pp reduction relative to tweet-level CV), and release the evaluation protocol for adoption by future work.

- **Signal specificity analysis and cross-dataset validation:** We provide controlled evidence that perspective signals encode epistemic information not captured by sentiment, stance, or uninformative auxiliary labels. Cross-dataset evaluation on PHEME establishes framework transferability, with variance reduction across unseen events providing evidence of stable generalization rather than overfitting to TACO-specific patterns.

2 Related Work

2.1 Conversational Argument Mining

Argument mining has evolved from identifying argumentative discourse units in formal texts (Stab and Gurevych, 2016; Palau and Moens, 2009) to analyzing informal conversational argumentation on social media (Habernal and Gurevych, 2017; Lawrence and Reed, 2019). Early approaches addressed monological structures in persuasive essays and legal documents using discourse markers and rhetorical relations (Feng and Hirst, 2011; Mayer et al., 2020). Adapting to Twitter's brevity and implicitness has required purpose-built datasets and models (Dusmanu et al., 2017; Bosc et al., 2016; Addawood and Bashir, 2016).

The TACO dataset (Feger and Dietze, 2024) provides fine-grained argumentative role labels within threaded Twitter conversations, enabling research on discourse-level argumentation patterns. State-of-the-art approaches on TACO employ BERTweet (Nguyen et al., 2020) and BERT (Devlin et al., 2019) fine-tuning on the four-class classification task. These approaches treat argumentative role assignment as a purely linguistic classification problem, conditioning only on tweet text and

optionally on thread structure. Our work demonstrates that this framing leaves exploitable signal on the table: cognitive perspective labels provide supervisory information that reshapes the representation space in ways that surface-level fine-tuning does not.

Evaluation methodology for conversational argument mining deserves scrutiny. The TACO benchmark and prior work following it employ tweet-level stratified cross-validation, which permits tweets from the same conversation to appear in both training and test folds. In a dataset where each tweet's argumentative role is partly determined by its discourse position within a thread, this constitutes a form of contextual leakage. We formalize this concern and address it through conversation-disjoint evaluation, providing both a corrected protocol and empirical quantification of the performance gap it introduces.

2.2 Cognitive Perspectives in Discourse Analysis

Epistemic cognition — how individuals understand the sources, certainty, and justification of knowledge — has been extensively studied in educational psychology and cognitive science (Hofer and Pintrich, 1997; Kuhn and Weinstock, 2002) but rarely operationalized in NLP. The distinction between epistemic beliefs (knowledge as objective and evidence-based) and experiential beliefs (knowledge derived from personal narrative) (King and Kitchener, 2004) maps directly onto observable patterns in argumentative text. Actively Open-Minded Thinking (Baron, 2008), a disposition toward evidence-based belief revision, constitutes a cognitively distinct reasoning mode with implications for argumentation quality.

Computational work on discourse perspectives has largely addressed surface-level properties: sentiment (Pang and Lee, 2008), stance (Mohammad et al., 2016), and media framing (Card et al., 2015). These approaches capture *what* users argue and *toward what* — their position and affect — but not *how* they epistemically construct their argument. Work on epistemic vigilance (SPERBER et al., 2010) in NLP has addressed fact-checking (Thorne et al., 2018)and claim verification (Wadden et al., 2022), but treats epistemology as a binary property (factual/non-factual) rather than as a multi-dimensional cognitive construct. Our taxonomy operationalizes epistemic cognition as a six-dimensional multi-label space, providing finer-grained distinctions that we demonstrate to be empirically informative for downstream classification.

LLM-based annotation has emerged as a productive paradigm for scaling theory-driven linguistic analysis (Ziems et al., 2024). Prior work has applied GPT-4 to annotate rhetorical moves in scientific writing and pragmatic functions in dialogue, but these efforts typically lack downstream task evaluation. We contribute both a structured annotation pipeline and empirical evidence linking annotation quality — as measured by signal specificity against controlled baselines — to downstream classification performance.

Our work occupies the intersection of three research threads: it brings epistemic cognition theory into argument mining (extending cognitive NLP beyond sentiment and stance), it contributes a rigorous evaluation methodology for conversational datasets (addressing a gap in how discourse context is handled in cross-validation), and it provides a transferable annotation framework validated across tasks and datasets. To our knowledge, this is the first work to integrate epistemic cognition theory into computational argument mining with quantitative downstream evaluation and controlled signal specificity analysis.

3 Methodology

We present a three-stage framework for enriching argumentative social media conversations with cognitively grounded perspective labels and evaluating their impact on discourse classification. A central design principle distinguishes our framework from prior work in conversational argument mining, we enforce **conversation-level data isolation** throughout the pipeline. Because tweets within the same thread share implicit discourse context, topic framing, register, and rhetorical positioning relative to prior turns, tweet-level cross-validation permits this context to leak across the train-test boundary, inflating performance estimates in ways that do not reflect generalization to unseen conversations. We eliminate this by construction at the evaluation design stage, as detailed in Section 3.2. Figure 1 provides an overview of the three-stage pipeline.

3.1 Perspective Annotation via Structured Prompting

Given conversational dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ where x_i is a tweet and y_i is a discourse label, we construct a multi-label perspective annotation

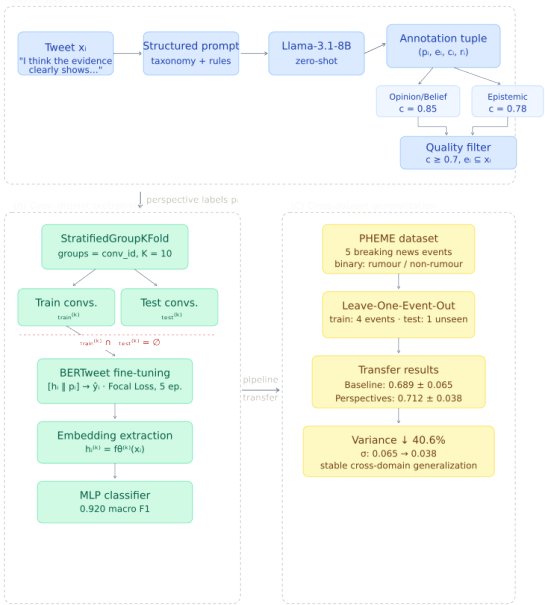

Figure 1: Overview of the perspective-enriched knowledge graph framework. (A) LLM-based perspective annotation with structured prompting. (B) Task-specific pretraining on TACO labels via 10-fold cross-validation. (C) Embedding extraction and node feature construction combining text, perspectives, and conversational metadata.

pipeline grounded in epistemic cognition theory. We define a taxonomy $\mathcal{T}$ of six cognitive dimensions (Baron, 2008) **Information Providing:** Active encoding and transfer of existing knowledge to another agent, motivated by a perceived informational need and involving memory retrieval, synthesis, and communication strategy. **Information Seeking:** Purposive effort to identify and acquire information in response to a perceived knowledge gap, involving query formulation, source evaluation, and iterative refinement of understanding. **Epistemic/Factual:** Emphasis on knowledge as objective, verifiable, and evidence-based, evaluated against established standards of justification to determine accuracy and reliability (Hofer and Pintrich, 1997). **Experiential:** Knowledge derived from and validated through direct personal experience, contrasting with abstract or detached claims and involving reflective learning through narrative (King and Kitchener, 2004). **Opinion/Belief:** Subjective mental representations held to be true without meeting rigorous epistemic standards, influenced by personal experience, social interaction, and cognitive bias. **Actively Open-Minded Thinking (AOT):** Metacognitive disposition to sincerely consider alternative perspectives, weigh evidence fairly, and override confirmation bias when forming conclusions (Baron, 2008).

For each tweet x_i, we construct a structured prompt containing precise category definitions with psychological grounding, multi-label assignment rules with explicit decision boundaries, and instructions for evidence extraction and confidence scoring. We query Llama-3.1-8B (Grattafiori et al., 2024) in a zero-shot setting. The output for each tweet is a tuple (p_i, e_i, c_i, r_i) where $p_i \subseteq \{1, \ldots, 6\}$ is the set of assigned perspective labels, e_i is a verbatim evidence span satisfying $e_i \subseteq x_i$, $c_i \in [0, 1]$ is a model confidence score, and r_i is a natural language rationale grounding the label assignment.

Quality Filtering We retain annotations satisfying $c_i \geq 0.7$ and enforce $e_i \subseteq x_i$ via exact substring matching, discarding approximately 22% of raw annotations. The evidence span requirement is a key quality control mechanism: a perspective label unsupported by any verifiable sub-string of the tweet text is flagged and removed, making hallucination directly detectable. Failed annotations are logged for review.

Annotation Independence Perspective labels are generated as a preprocessing step, prior to any cross-validation splitting and without access to ground-truth discourse labels. This ensures that the annotation process cannot encode target label information, a concern we address empirically through auxiliary signal comparison in Section 3.3

Table 1 shows the resulting perspective distribution across argumentative roles TACO. Perspective profiles are class-discriminative: Reason tweets exhibit elevated Epistemic (12.6% vs. 6.2% overall) and Experiential (17.7% vs. 8.7% overall) reasoning alongside dominant Opinion/Belief (88.0%), reflecting their evidential and experiential grounding. Notification tweets are dominated by Information Providing (74.8%) with notably lower Opinion content (42.4%), consistent with their broadcast function. Statement tweets show mixed profiles, high Opinion (86.3%) but low Information Providing (18.7%), a signature that explains their historical classification difficulty, lacking both evidential grounding of Reason tweets and the informational clarity of Notifications, Statement tweets occupy an ambiguous epistemic position that surface-level features struggle to resolve.

Perspective	Reason (n=581)	Statement (n=284)	Notification (n=500)	None (n=369)
Information Providing	60.9%	18.7%	74.8%	18.4%
Information Seeking	7.6%	10.9%	3.0%	6.8%
Epistemic/Factual	12.6%	1.8%	6.8%	0.3%
Experiential	17.7%	6.7%	5.8%	1.6%
Opinion/Belief	88.0%	86.3%	42.4%	62.6%
AOT	2.2%	2.1%	0.0%	0.3%
Avg perspectives/tweet	1.89	1.26	1.33	0.90

Table 1: Distribution of cognitive perspectives (%) across argumentative role categories in the perspective-enriched TACO dataset. Each column is independent, values need not sum to 100 as tweets may carry multiple perspective labels.

3.2 Conversation-Disjoint Representation Learning

Evaluation protocol We employ **StratifiedGroupKFold** cross-validation ($K = 10$, `random_state=123456789`) with conversation IDs as the grouping variable. Formally, let $\mathcal{C}$ denote the set of all conversation IDs in $\mathcal{D}$. For fold k, we divide $\rfloor$ into disjoint subsets:

$$\mathcal{C}_{\text{train}}^{(k)} \cap \mathcal{C}_{\text{test}}^{(k)} = \emptyset \quad (1)$$

All tweets x_i whose conversation ID belongs to $\mathcal{C}_{\text{test}}^{(k)}$ constitute the test set for fold k; no tweet from these conversations appear in any training fold. Crucially, stratification is applied at the *tweet level* rather than the conversation level, ensuring that each fold's test split reflects the full class distribution of the dataset. This is non-trivial: conversation-mode stratification, grouping conversations by their dominant class label, does not guarantee class coverage in every fold, producing degenerate splits where minority classes are absent from the test sets. StratifiedGroupKFold simultaneously enforces group disjointness and class stratification, yielding consistent minority class representation across folds (28-29 Statement test instances per fold across all 10 folds in TACO)

Model architecture We fine-tune BERTweet (Nguyen et al., 2020), a RoBERTa model pretrained on 850M English tweets, as the backbone encoder. For fold k, the classifier concatenates the `[CLS]` token embedding with the binary perspective feature vector prior to classification:

$$\hat{y}_i = \text{softmax}(W_2\, \sigma(W_1\, [h_i \,\|\, p_i] + b_1) + b_2) \quad (2)$$

where $h_i \in R^{768}$ is the `[CLS]` embedding from BERTweet's penultimate layer, $p_i \in \{0,1\}^6$ is the binary perspective vector, σ is ReLU activation, and $\|$ denotes concatenation. Perspective features act as supervisory signals during fine-tuning: gradients flow through the full network, enabling the encoder to internalize cognitive-epistemic distinctions into its representation space rather than appending them as post-hoc features. The baseline variant omits p_i, fine-tuning BERTweet on discourse labels only.

Training uses AdamW (Loshchilov and Hutter, 2019) with learning rate 4×10^{-4}, weight decay 10^{-2}, batch size 8, and 5 epochs. To address class imbalance (Reason: 33.5%, Statement: 16.4%, Notification: 28.8%, None: 21.3%), we employ Focal Loss (Lin et al., 2018):

$$\mathcal{L}_{\text{focal}} = -\alpha_y (1 - p_y)^{\gamma} \log p_y \quad (3)$$

with $\gamma = 2$ and per-fold inverse-frequency class weights $\alpha_c = N/(C \cdot N_c^{(k)})$, where $N_c^{(k)}$ counts training instances of class c in fold k. Recomputing weights per fold ensures that no test-fold class distribution information influences training.

The learning rate of 4×10^{-4} is higher than typical full fine-tuning regimes (2–5 $\times 10^{-5}$) because we fine-tune with a small batch size (8) over 5 epochs on a small dataset (1,560 training tweets per fold), where lower learning rates produce insufficient parameter updates within the available training budget. No learning rate warm-up or scheduling is applied, we observed stable training loss curves across all 10 folds without instability, consistent with the small effective step count ($\sim$975 steps per fold).

Fold-aware embedding extraction After fine-tuning fold k's model $f_\theta^{(k)}$, we extract the fused representation for each tweet:

$$h_i^{(k)} = \text{ReLU}(W_1\, [h_i^{\text{CLS}} \,\|\, p_i] + b_1) \in R^{768} \quad (4)$$

where $h_i^{\text{CLS}} \in R^{768}$ is the `[CLS]` token embedding from BERTweet's penultimate layer and $p_i \in \{0,1\}^6$ is the same binary perspective vector used during fine-tuning. Perspective features are thus present at both fine-tuning and embedding extraction, ensuring that the extracted representation reflects the encoder's perspective-conditioned internal state. The downstream MLP is trained strictly on training-fold embeddings and evaluated on test-fold embeddings. Concretely, for fold k: (i) $f_\theta^{(k)}$ is trained only on conversations in $\mathcal{C}_{\text{train}}^{(k)}$;

(ii) embeddings $\{h_i^{(k)}\}$ are extracted for *all* tweets using $f_\theta^{(k)}$, but the MLP classifier is fit only on $\{h_i^{(k)} : \mathrm{conv}(i) \in \mathcal{C}_{\mathrm{train}}^{(k)}\}$; (iii) evaluation uses only $\{h_i^{(k)} : \mathrm{conv}(i) \in \mathcal{C}_{\mathrm{test}}^{(k)}\}$, which come from a model that never saw their labels or conversational context. This two-stage design cleanly isolates the degree to which perspective supervision reshapes the representation space from the classification head's capacity.

3.3 Auxiliary Signal Specificity Analysis

A necessary condition for claiming that epistemic perspective signals are specifically informative is demonstrating that equivalent gains cannot be obtained from generic auxiliary labels. We design a controlled comparison holding all experimental factors constant, architecture, hyperparameters, conversation-disjoint splits, while varying t=only the supervision signal type.

Sentiment Three-class sentiment labels (negative/positive/neutral) are generated using `cardiffnlp/twitter-roberta-base-sentiment-latest` (Barbieri et al., 2022), a Twitter-specific sentiment model. Sentiment captures the affective tone of a tweet, how a user *feels* about a topic, rather than epistemic mode through which they construct their argument.

Stance Three-class stance labels (against/favor/neutral) are generated using `cardiffnlp/twitter-roberta-base-stance-semeval2016` (Mohammad et al., 2016). Stance captures positional orientation, what a user thinks about a topic, rather than the cognitive process underlying their claim construction.

Random binary features Random binary vectors are sampled with dimensionality six and Bernoulli density 0.225, matching the average per-label prevalence of perspective annotations. This baseline controls for feature dimensionality and sparsity, isolating the contribution of semantic context from that of feature shape.

All three auxiliary signals are encoded identically to perspective features: as fixed-width vectors concatenated with BERTweet `[CLS]` embeddings, with the encoder fine-tuned end-to-end. If perspective-enriched models outperform all three controls under this controlled comparison, this constitutes evidence that epistemic perspective signals are task-specifically informative beyond what generic discourse properties provide.

3.4 Cross-Dataset Generalization on PHEME

To evaluate framework transferability, we apply the annotation pipeline to PHEME (Zubiaga et al., 2016), a dataset of TWITTER source tweets labeled as rumour or non-rumour across five breaking news events. The theoretical motivation for this transfer is grounded in shared epistemic structure, rumour engagement in breaking news contexts activates the same reasoning modes our taxonomy captures. Users spreading or correcting claims engage Information Providing, users citing verification evidence engage Epistemic/Factual reasoning, users asserting veracity without evidence engage Opinion/Belief. The perspective taxonomy therefore encodes reasoning patterns that are not task-specific to argumentative role classification but reflect domain-general epistemic behavior in social media discourse.

We annotate all PHEME source tweets using the same Llama-3.1-8B pipeline with identical prompting and filtering. Classification targets the binary rumour/non-rumour label. We evaluate under **Leave-One-Event-Out (LOGO)** cross-validation, training on four events and testing on the held-out fifth across all five event permutation. For each LOGO fold, we train both a BERTweet baseline and a perspective-enriched model under identical conditions. Performance gap and inter-event variance jointly quantify the contribution of perspective signals to stable cross-domain generalization.

4 Evaluation and Results

4.1 Dataset

TACO The Twitter Argument from COnversation dataset (Feger and Dietze, 2024) contains 1,814 tweets across 200 threaded conversations spanning six heterogeneous topics (Brexit, Game of thrones, Twitter Takeover, Abortion, Squid Game, LOTR), annotated with four argumentative role labels: Reason, Statement, Notification, and None. After removing Undecided instances, the dataset comprises 1,734 tweets (Reason: 33.5%, Notification: 28.8%, None: 21.3%, Statement: 16.4%). We extend TACO by enriching each tweet with perspective annotations through the pipeline described in Section 3.1.

PHEME The PHEME dataset (Zubiaga et al., 2016) contains Twitter conversations collected during five breaking news events,Charlie Hebdo (2,079 tweets), Ferguson (1,143 tweets), Germanwings

Model	Mac. F1	Wt. F1	Std
Direct Classification			
BERTweet Baseline	0.7083	0.7264	∼0.05
+ Random features	0.6784	0.6886	0.0568
+ Sentiment	0.6662	0.7147	0.1426
+ Stance	0.6996	0.7200	0.0496
+ Perspectives	0.7005	0.7179	∼0.05
Embedding-Based Classification			
MLP on Vanilla Embeddings	0.7976	0.8147	0.0334
MLP on Perspective Embeds	**0.9217**	**0.9373**	**∼0.0252**

Table 2: Performance on TACO under conversation-disjoint StratifiedGroupKFold evaluation ($K = 10$). All direct classification models share BERTweet architecture and hyperparameters; only the auxiliary signal type varies. **Bold** indicates best result.

Model	Reason	Stmt.	Notif.	None
BERTweet Baseline	0.751	0.5331	0.7637	0.7847
BERTweet + Perspectives	0.7332	0.5269	0.7656	0.7762
MLP (Vanilla Embs.)	0.8461	0.6539	0.8532	0.8371
MLP (Persp. Embs.)	0.9481	0.8835	0.9439	0.9116

Table 3: Per-class F1 under conversation-disjoint StratifiedGroupKFold evaluation (mean ± std across 10 folds). Perspective-enriched embeddings achieve strong performance across all classes, with the largest absolute gain for Statement

Crash (469 tweets), Ottawa Shooting (890 tweets), and Sydney Siege (1,221 tweets). Each source tweet is labeled as rumour or non-rumour, with class balance varying across events (49-78% non-rumour). We annotate all source tweets using our cognitive perspective taxonomy pipeline.

4.2 Main Results on TACO

Table 2 presents results under conversation-disjoint StratifiedGroupKFold evaluation across all model configurations. We report macro F1 as the primary metric given class imbalance, with weighted F1 and standard deviation across folds.

Three findings emerge from Table 2:

Signal specificity: Sentiment (0.666) and stance (0.699) both fall below the BERTweet baseline (0.7083), as does random baseline (0.678). This result is substantively important since it establishes that the benefit of auxiliary supervision is not a generic artifact of multi-label feature injection but is specific to the epistemic content of perspective labels. Sentiment and stance are well-established discourse signals with strong Twitter-specific models, their failure to match the baseline, let alone the perspective-enriched model, rules out the hypothesis that any auxiliary signal of comparable quality would produce equivalent gains.

Supervisory mechanism: In direct classification, perspective features match but do not substantially exceed the baseline (0.7083 vs. 0.7005). The performance discontinuity appears at the embedding stage: MLP on perspective-enriched embeddings achieves 0.922, a 12.2pp gain over vanilla embeddings (0.798). This pattern is consistent with perspectives acting as *supervisory signals that reshape the representation space* during fine-tuning rather than as additive features at classification time. The encoder internalizes cognitive-epistemic distinctions during training, the downstream MLP merely reads them off.

Evaluation robustness: The conversation-disjoint protocol reduces performance relative to tweet-level cross-validation by 1.2pp (0.922 vs. 0.934; tweet-level results reported in Appendix C for benchmark comparability). This modest reduction confirms that the gains are not primarily attributable to conversational context leakage and that the framework generalizes to conversations unseen during training.

Table 3 disaggregates performance by argumentative role.

The per-class breakdown revels that the perspective enrichment provides the largest absolute gains for Statement, the minority class with the most ambiguous surface form. This is consistent with the perspective distribution analysis in Table 1: Statement tweets exhibit high Opinion/Belief (86.3%) but low information providing (18.7%), a signature that distinguishes them from superficially similar Reason tweets (which also exhibit high opinion but additionally show elevated Epistemic and Experiential reasoning). The encoder trained with perspective supervision learns to exploit this distinguishing signal, the baseline encoder, conditioning only on text, cannot reliably separate these classes.

The 12.2pp improvement from vanilla to perspective-enriched embeddings (0.798 → 0.920 macro F1) on a dataset of 1,734 merits careful examination. Several analyses help establish that the gain reflects meaningful signal rather than artifacts of the training setup. First, the conversation-disjoint protocol ensures that no tweet in any test fold shares conversational context with any training tweet, eliminating the most plausible source of inflation. Second, the gains persist under the stricter

Test Event	Baseline	Perspectives
Charlie Hebdo	0.7595	0.7581
Ferguson	0.5899	0.6611
Germanwings	0.6652	0.7408
Ottawa Shooting	0.7123	0.6970
Sydney Siege	0.7168	0.7047
Mean $\pm$ Std	0.688 ± 0.064	$\mathbf{0.712 \pm 0.038}$

Table 4: Leave-One-Event-Out macro F1 on PHEME rumour detection. Each row trains on four events and tests on the held-out fifth.

PHEME LOGO evaluation (Section 4.3), where the model is tested on entirely unseen news events, an evaluation setting that would immediately expose overfitting to TACO-specific patterns. Third, the ablation in Table **??** shows that concatenating perspective features to *vanilla* embeddings yields no improvement (0.816 macro F1), ruling out the features themselves as the source; the gain is specifically attributable to perspective-supervised *fine-tuning* reshaping the representation space. Fourth, the auxiliary signal comparison establishes that equivalent gains cannot be obtained from other auxiliary signals of comparable quality, including those generated by strong Twitter-specific models (sentiment, stance). Taken together, these controls provide convergent evidence that the gains reflect genuine signal in the perspective taxonomy rather than methodological artifact.

4.3 Cross-Dataset Generalization on PHEME

Table 4 presents Leave-One-Event-Out results on PHEME. We report per-event results along with aggregate statistics to make transparent whether gains are consistent across events or driven by individual held-out conditions.

Two findings characterize the PHEME results:

Cross-event pattern: Perspective enrichment produces large improvements on two events (Ferguson: $+7.1$pp, Germanwings: $+7.6$pp) while performing comparably or slightly below baseline on three others (Charlie Hebdo: -0.1pp, Ottawa: -1.5pp, Sydney: -1.2pp). The mean gain of $+2.4$pp thus reflects a heterogeneous per-event pattern rather than uniform improvement. However, the aggregate variance reduction tells the more informative story, the baseline's high inter-event variance ($\sigma = 0.064$) reflects brittleness to the choice of test event, particularly the low-performing Ferguson fold (0.590). Perspective enrichment raises this floor substantially (Ferguson: 0.661) while maintaining comparable performance on high-performing events, compressing the inter-event distribution ($\sigma = 0.038$, -40.6%). This variance reduction is the primary generalization signal, it demonstrates that perspective-enriched representations are more robust to domain shift across unseen breaking news events, even when mean improvement is modest.

The theoretical basis for this transfer is grounded in shared epistemic structure across tasks. rumour engagement in breaking news activates the same reasoning modes the cognitive perspective taxonomy encodes: Information providing captures claim propagation, Epistemic/Factual captures verification behavior, and Opinion/Belief captures unverified assertion. The framework's effectiveness on PHEME thus reflects the domain-generality of epistemic cognition as a theoretical framework, not incidental overlap between TACO and PHEME.

5 Conclusion

We introduce the framework for mining cognitive perspectives in social media argumentation, grounding argument mining in epistemic cognition theory and demonstrating that cognitively motivated perspective labels function as effective supervisory signals for argumentative role classification. Under conversation-disjoint evaluation, which corrects for a methodological gap in prior conversational argument mining work, perspective-enriched embeddings achieve 0.92 macro F1 on TACO, outperforming text-only baseline by 12.2pp. A controlled auxiliary signal comparison establishes signal specificity; sentiment, stance, and random baselines under identical conditions fall below the unenriched baseline, ruling out generic auxiliary supervision as an explanation for perspective gains. Cross-dataset evaluation on PHEME rumour detection demonstrates cross-task transferability, with perspectives improving mean macro F1 by 2.4pp and reducing cross-event variance by 40.6% under Leave-One-Event-Out evaluation.

Beyond the empirical results, this work makes a methodological contribution to conversational NLP, conversation-level disjointness in cross-validation is a validity concern for any task where the target label is partly determined by discourse position within a thread. The StratifiedGroupKFold protocol we implement and quantify is directly applicable to other conversational datasets and evaluation settings.

Limitations

Evaluation scope Our conversation-disjoint protocol addresses within-dataset contextual leakage by ensuring no conversational thread bridges train and test split. Topic-disjoint evaluation, holding out all conversations from a given debate topic, would provide stricter test of generalization across discourse domains and is a natural next-step given TACO's six-topic structure. We also note that while StratifiedGroupKFold ensures consistent minority class coverage across folds, the small absolute size of the dataset (1,734 tweets, 284 statement instances) limits the statistical power of per-fold estimates, fold-level standard deviations reflect genuine evaluation variance rather than model instability.

Annotation reliability Perspective labels are generated by a single model (Llama-3.1-8B) with confidence filtering at $c \geq 0.7$. Multi-model agreement analysis reported in Appendix **??** finds Fair cross-model agreement (mean Cohen's κ: 0.22–0.32 across model pairs), with higher agreement between larger models (Llama-70B vs. Mistral-7B: $\kappa = 0.32$) and on high-prevalence labels (Information Seeking: $\kappa = 0.51$, Opinion/Belief: $\kappa = 0.47$ for the 70B/Mistral pair). Lower agreement on AOT ($\kappa = 0.09$–0.16) reflects its rarity (1–6% prevalence) rather than definitional ambiguity: kappa is systematically unreliable for rare labels due to low base rate variance. These results motivate ensemble annotation strategies for future work, particularly for low-prevalence perspective categories. We also note that Fair kappa between model families is consistent with known calibration differences across LLM scales, the 8B model assigns Information Providing to 28.7% of tweets versus 49.0% for the 70B model, rather than indicating taxonomy ambiguity.

Thread-agnostic annotation Our annotation pipeline processes each tweet independently, without access to the parent tweet or conversational thread. This is a deliberate design choice that simplifies deployment and ensures annotation portability across datasets with varying thread structures. However, some perspective labels, particularly Information Seeking and AOT, are more naturally interpreted relative to a conversational context. Thread-aware annotation models, conditioning on the full reply chain, may improve precision for these labels while introducing dependencies on thread availability.

Language and platform scope. All experiments use English Twitter data. The perspective taxonomy is grounded in psychological theory that is not language-specific, but the annotation pipeline and downstream models are calibrated to English Twitter conventions. Extension to other languages would require multilingual backbone models and potentially language-specific prompt adaptation. Platform-specific linguistic norms, Twitter's brevity constraints, hashtag conventions, and threading semantics, may also limit direct transferability to other social media platforms without recalibration.

References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Jonathan Baron. 2008. Actively open-minded thinking. *Thinking and deciding*, 188:199–232.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Marc Feger and Stefan Dietze. 2024. Taco – twitter arguments from conversations. *Preprint*, arXiv:2404.00406.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Barbara Hofer and Paul Pintrich. 1997. The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, 67.

Patricia King and Karen Kitchener. 2004. Reflective judgment: Theory and research on the development of epistemic assumptions through adulthood. *Educational Psychologist - EDUC PSYCHOL*, 39:5–18.

Deanna Kuhn and Michael Weinstock. 2002. What is epistemological thinking and why does it matter? *Personal epistemology: The psychology of beliefs about knowledge and knowing*.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. *Preprint*, arXiv:1801.07606.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *European Conference on Artificial Intelligence*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

DAN SPERBER, FABRICE CLÉMENT, CHRISTOPHE HEINTZ, OLIVIER MASCARO, HUGO MERCIER, GLORIA ORIGGI, and DEIRDRE WILSON. 2010. Epistemic vigilance. *Mind & Language*, 25(4):359–393.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *Preprint*, arXiv:1604.07370.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, and 1 others. 2020. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. 2016. PHEME dataset of rumours and non-rumours.

A Appendix

B PHEME Perspective Distribution

Table 5 reports the distribution of perspective labels across PHEME source tweets, disaggregated by rumour/non-rumour label. These distributions motivated the theoretical connection between the perspective taxonomy and rumour detection discussed in Section 3.4.

2*Perspective	2*Overall	By Label	
		Rumour	Non-rumour
Info. Providing	24.3%	25.1%	23.9%
Info. Seeking	8.0%	9.5%	7.4%
Epistemic/Factual	8.1%	9.6%	7.4%
Experiential	3.8%	3.2%	4.0%
Opinion/Belief	49.4%	46.2%	50.8%
AOT	2.5%	2.2%	2.7%

Table 5: Perspective label distribution in PHEME (102,348 total annotations across 5,962 source tweets). Rumour tweets show elevated Information Seeking (+2.1pp) and Epistemic/Factual (+2.2pp) reasoning relative to non-rumours, consistent with the epistemic uncertainty characteristic of unverified claims. Non-rumours show higher Opinion/Belief (+4.6pp), reflecting established journalistic commentary framing.

The perspective distribution in PHEME differs substantially from TACO (Table 1), reflecting the domain shift from structured Twitter debates to breaking news commentary. Opinion/Belief dominates in PHEME (49.4%) at similar rates to TACO's overall prevalence (50.4%), while Epistemic/Factual is notably lower (8.1% vs. 12.6% for Reason tweets in TACO). The pattern of elevated Information Seeking and Epistemic/Factual in rumour tweets relative to non-rumours is theoretically interpretable, unverified claims generate information-seeking responses and prompt epistemic evaluation, whereas confirmed or clearly non-rumour content elicits commentary framing (Opinion/Belief).

C Tweet-Level Cross-Validation Results

Table 6 reports results under tweet-level startified cross-validation for direct comparison with prior TACO work (Feger and Dietze, 2024). These results are *not* our primary evaluations, we report them solely to enable benchmark comparability. As discussed in Section 3.2, tweet-level CV permits tweets from the same conversation to appear in both training and test folds, constituting a form of contextual leakage that inflates performance estimates.

Model	Acc.	Mac. F1	Wt. F1
Text-Based (Direct Classification)			
BERTweet Baseline	0.7336	0.7161	0.7330
BERTweet + Perspectives	0.7388	0.7247	0.7408
Embedding-Based			
MLP (Vanilla Embeds.)	$0.8322_{\pm.022}$	$0.8158_{\pm.029}$	$0.8308_{\pm.024}$
MLP (Persp. Embeds.)	$\mathbf{0.9394}_{\pm.010}$	$\mathbf{0.9339}_{\pm.012}$	$\mathbf{0.9392}_{\pm.010}$
Graph-Based (Perspective Embeddings)			
GraphSAGE (3-layer)	$0.9337_{\pm.011}$	$0.9267_{\pm.012}$	$0.9334_{\pm.011}$
GAT (4 heads)	$0.9291_{\pm.012}$	$0.9228_{\pm.014}$	$0.9289_{\pm.012}$
GCN (3-layer)	$0.9198_{\pm.009}$	$0.9140_{\pm.009}$	$0.9199_{\pm.009}$

Table 6: Tweet-level stratified cross-validation results on TACO, reported for benchmark comparability with prior work. These results should not be compared directly with Table 2, which uses conversation-disjoint evaluation. GNN results are included here as they were evaluated under the same tweet-level protocol; under conversation-disjoint evaluation, GNNs are omitted from the main paper as they do not add to the primary claims.

The tweet-level results reveal that GNN architectures consistently underperform the MLP classifier when operating on perspective-enriched node features (GraphSAGE: 0.927, GAT: 0.923, GCN: 0.914 vs. MLP: 0.934). We attribute this to two factors. First, TACO's conversational structure produces a small (N=1,734), densely connected graph with three edge types, a topology associated with rapid feature homogenization through over-smoothing (Li et al., 2018). Second, perspective-aware fine-tuning produces representations that already encode discourse-structural signals, explicit neighborhood aggregation via GNN propagation introduces noise rather than additional information. This finding is consistent with the recent theoretical work showing that strong node features can render graph structure redundant for classification (Zhu et al., 2020).

D Full Ablation Results with Standard Deviations

Setting	**Acc.**	**Mac. F1**	**Wt. F1**
Embeddings only	$0.832_{\pm.022}$	$0.816_{\pm.029}$	$0.831_{\pm.024}$
+ Perspectives	$0.826_{\pm.026}$	$0.808_{\pm.033}$	$0.824_{\pm.029}$
+ Topic	$0.829_{\pm.017}$	$0.814_{\pm.024}$	$0.828_{\pm.019}$
+ Depth	$0.827_{\pm.024}$	$0.811_{\pm.032}$	$0.826_{\pm.026}$
+ Confidence	$0.831_{\pm.023}$	$0.815_{\pm.029}$	$0.830_{\pm.025}$
Persp.-aware Embeds.	$\mathbf{0.939}_{\pm.010}$	$\mathbf{0.934}_{\pm.012}$	$\mathbf{0.939}_{\pm.010}$
Perspectives only	$0.419_{\pm.026}$	$0.280_{\pm.022}$	$0.340_{\pm.024}$
Topic only	$0.395_{\pm.019}$	$0.273_{\pm.027}$	$0.305_{\pm.030}$
Depth only	$0.369_{\pm.026}$	$0.230_{\pm.017}$	$0.286_{\pm.021}$
Confidence only	$0.335_{\pm.002}$	$0.126_{\pm.001}$	$0.168_{\pm.002}$

Table 7: Full ablation results with standard deviations across 10 folds. Upper block: feature concatenation ablations using vanilla BERTweet embeddings. Perspective-aware embeddings shown for reference. Lower block: individual features used as sole input.