Codebook-Injected Dialogue Segmentation for Multi-Utterance Constructs Annotation: LLM-Assisted and Gold-Label-Free Evaluation

Anonymous ACL submission

Abstract

Dialogue Act (DA) annotation typically treats communicative or pedagogical intent as localized to individual utterances or turns. This leads annotators to agree on the underlying action while disagreeing on segment boundaries, reducing apparent reliability. We propose *codebook-injected segmentation*, which conditions boundary decisions on downstream annotation criteria, and evaluate LLM-based segmenters against standard and retrieval-augmented baselines. To assess these without gold labels, we introduce evaluation metrics for span consistency, distinctiveness, and human-AI distributional agreement. We found DA-awareness produces segments that are internally more consistent than text-only baselines. While LLMs excel at creating construct-consistent spans, coherence-based baselines remain superier at detecting global shifts in dialogue flow. Across two datasets, no single segmenter dominates: improvements in within-segment coherence frequently trade off against boundary distinctiveness and human–AI distributional agreement. These results highlight segmentation as a consequential design choice that should be optimized for downstream objectives rather than a single performance score.

1 Introduction

High-quality labeled data is a major bottleneck for building reliable language technologies in high-stakes settings, where small annotation decisions can shift both empirical conclusions and model behavior. This has motivated extensive work on how to scale labeling while maintaining quality, including agreement measurement and pipeline design (Artstein and Poesio, 2008), the use of non-expert annotators to expand supervision (Snow et al., 2008), and crowdsourcing protocols for collecting human judgments (Kittur et al., 2008). However, the usefulness of a labeled dataset ultimately depends not only on reliability between annotators, but also on whether the labels faithfully capture the target constructs and can be reproduced at scale. This is particularly important for dialogue act (DA) annotation, where annotators may assign the same construct to adjacent utterances, creating boundary-driven disagreement and inflating inter-annotator disagreement (Ostyakova et al., 2022; Midgley, 2009).

DA annotation tasks encounter several design challenges, including separating semantic from pragmatic meaning, handling metadiscourse, and choosing a tagset that is both expressive and learnable (Verdonik, 2023). These challenges are not only about label choice; they also reflect a unit-of-analysis problem where coders must decide both *what* the label is and *where* it starts and ends. Disagreement can therefore reflect boundary misalignment, not just conceptual mismatch. Work on agreement for segmentation and "unitizing" continuous records similarly emphasizes that boundary placement is itself a major source of unreliability and should be modeled explicitly (Krippendorff et al., 2016; Mathet et al., 2015).

A practical response is to introduce an intermediate span level and treat segmentation as an explicit first step in the annotation workflow. Segmentation has a long history in NLP, particularly in topic and discourse analysis, and recent dialogue-specific methods improve segmentation robustness by modeling local coherence or learning representations tailored to boundary detection (Xing and Carenini, 2021; Xing et al., 2020; Gao et al., 2023). At the same time, the expansion of supervised segmentation resources has enabled stronger learned models and more principled benchmarking of boundary definitions (Jiang et al., 2023). More recent work has explored the use of large language models (LLMs) for discourse analysis tasks, including dialogue topic segmentation, often through structured prompting that externalizes the model's boundary decisions into an interpretable output format (Fan et al., 2024; Das et al., 2024). In parallel, human-LLM collaborative annotation pipelines increasingly emphasize verification and oversight, acknowledging that LLM-generated annotations scale effectively only when humans can audit and intervene on intermediate decisions, including segmentation boundaries (Kim et al., 2024).

This paper addresses the unit-of-analysis problem by using examples from teaching dialogue, in which the goal is to identify what instructional methods an instructor uses. These types of dialogue pose a segmentation challenge because instructors' intent is contextual and often unfolds over multiple turns (Na and Feng, 2025; Hennessy et al., 2016). In this context, we define **tutoring moves as pedagogical dialogue acts**, functional units of instructional intent realized through a specific annotation codebook. Because these tutoring moves and instructional phases are realized as short spans rather than single utterances, boundary placement becomes an essential part of the annotation judgment. We therefore decouple boundary placement from label assignment, treating boundary decisions as an explicit, separable component of the annotation process so that multi-utterance constructs can be labeled more faithfully. The main contributions of our work are as follows:

1. We propose a codebook-injected (DA-aware) segmentation process that introduces an intermediate span level for annotation, explicitly decoupling boundary placement from label assignment so that multi-utterance constructs can be labeled more faithfully.

2. We evaluate LLM-based segmenters with both generic topic-shift prompting and DA label guidance, and benchmark them against established NLP dialogue segmentation algorithms with a retrieval-augmented variant that injects DA cues from semantically matched examples and DA labels.

3. We introduce gold-label-free evaluation criteria for segmentation in the absence of reference labels, capturing within-segment consistency, adjacent segment distinctiveness, and segment-level distributional agreement.

2 Related Work

2.1 Annotating Dialogue Acts with LLMs

Researchers increasingly utilize LLMs to scale annotation, with performance relative to human experts varying by task complexity and workflow design (Su and Ye, 2025; Yu et al., 2024). While LLMs excel in constrained settings, complex discourse annotation often requires structured pipelines to handle real-world taxonomies and multi-label cases (Ostyakova et al., 2023). Recent shifts treat LLM annotation as a *process design* problem, improving robustness through multi-model ensembling and orchestration (Na and Feng, 2025; Ahtisham et al., 2025). For instance, frameworks like EduDCM improve reliability by decomposing constructs into specific components, such as acts and events, while integrating consistency checks across multiple LLMs (Qi et al., 2024).

However, multiple studies have documented persistent limitations of LLMs for DA annotation. In particular, LLMs can struggle with DA classification in multi-party settings such as classroom or group discussions with multiple speakers and overlapping interactional threads (Qamar et al., 2025). Moreover, the complexity of goal-oriented dialogue annotation schemas (e.g., dialogue state tracking and dialogue acts) can exceed the capabilities of current models, requiring substantial human oversight (Labruna et al., 2023). These analyses also emphasize concrete failure modes that matter for data quality, including hallucinated labels and instruction-following breakdowns that can degrade annotation quality relative to human coders (Labruna et al., 2023). In response, some work has explored adapting models or schemas to better reflect pragmatic intent, for example, via intent-focused approaches tied to education taxonomies (Petukhova and Kochmar, 2025), or by proposing dialogue schemas intended to be more usable for novice annotators while capturing both semantic and syntactic structure (Duran et al., 2022).

To mitigate these issues in practice, recent work increasingly emphasizes human-in-the-loop or reliability-oriented pipelines around LLM labeling. For example, studies propose workflows that integrate verification and adjudication steps to increase reliability in semi-automated annotation (Tavakoli and Zamani, 2025; Shah et al., 2025). At the same time, not all prompting strategies help. Mohammadi et al. (2025) find that persona prompting does not improve performance on fairness-related annotation tasks (e.g., sexism). Overall, these findings suggest that using LLMs to annotate DAs at the utterance level remains challenging, and that improvements often come from workflow design, decomposition, and oversight rather than prompting alone.

2.2 Dialogue Segmentation

Dialogue segmentation partitions interactions into contiguous, coherent segments based on topic, task, or discourse function, with boundaries signaling meaningful shifts. Early research on monologic text utilized lexical cohesion through methods like TextTiling (Hearst, 1997) and C99 (Choi, 2000) to detect local drops in lexical similarity, while later Bayesian and probabilistic approaches modeled segments via latent topic distributions to infer boundaries without labeled data (Eisenstein and Barzilay, 2008; Misra et al., 2011). Dialogic text introduces complexities such as lexical sparsity and pragmatic shifts, leading researchers to integrate conversational structure with lexical chains in methods like LCseg (Galley et al., 2003) or global objectives such as minimum cuts (Malioutov and Barzilay, 2006). Recently, the field has transitioned toward neural and representation-based methods that capture long-range dependencies (Xing et al., 2020), utilizing unsupervised techniques to detect coherence shifts between utterances (Gao et al., 2023) or framing the task as supervised boundary prediction using section-level labeled corpora (Koshorek et al., 2018).

Segmentation is most commonly framed as a topical or discourse coherence problem, evaluated against headings, agenda items, or coarse task boundaries. Far less work treats segmentation as a deliberate pre-annotation step intended to improve the reliability and interpretability of downstream labels. Although recent work compares LLM-based segmenters to traditional methods for segmenting open-ended conversations into problem-based units aligned with predefined materials (Wang et al., 2024), connections between segmentation choices and annotation reliability remain underexplored. This gap is especially salient in educational dialogue, where annotation targets reflect interactional strategies or pedagogical intent rather than topic alone, and boundaries may appear without strong lexical cues.

Practically, when annotating a construct of interest that appears at the segment level, enforcing strictly local labeling can inflate apparent disagreement: annotators may share an interpretation of a multi-turn span but anchor its onset/offset differently. Utterance-aligned agreement measures can then penalize near matches while obscuring span-level convergence (Artstein and Poesio, 2008). We therefore frame **segmentation as a boundary setting problem** aimed at reducing ambiguity and supporting reliable segment-level labeling.

3 Data and Context

To evaluate various segmentation methods, we utilize two independent educational dialogue datasets from distinct contexts, each expert-annotated with instructional move taxonomies that represent diverse pedagogical perspectives: the open-sourced TalkMoves dataset of authentic classroom discourse (Suresh et al., 2022), and a custom dataset of tutoring chat transcripts annotated using the Classroom Assessment Scoring System (CLASS) framework (Pianta et al., 2008). Educational dialogue annotation provides an ideal test case for segmentation because instructional 'moves' (i.e. domain-specific dialogue acts such as revoicing or scaffolding) often exist at a meso-level between a single utterance and a full session; for instance, providing an explanation may span multiple utterances and conversational turns (e.g., consider a student interjecting "ok", "yes", or "why is that" while a tutor is giving an explanation). These datasets, described in detail below and in Table 1, capture educational interaction and are grounded in distinct conceptualizations of instructional quality, enabling evaluation across both divergent dialog settings and complementary theoretical constructs.

Table 1: Overview of datasets and label spaces.

Feature	**TalkMoves**	**CLASS-Annotated**
Context	K–12 Math Lessons (Suresh et al., 2022)	Chat-based secondary school math tutoring with annotations of instructional support dimensions (Pianta et al., 2008)
Total Sessions	63	30
Total Utterances	31,263	1,881
Avg. Utterances	496	63
Labels	Keeping Everyone Together; Getting students to relate to an other's ideas; Pressing for Accuracy; Pressing for reasoning; Revoicing; Restating	Scaffolding; Building on Student Responses; Feedback Loop; Encouragement

3.1 TalkMoves

The TalkMoves dataset[1] is an open-source collection of mathematics classroom transcripts from authentic K-12 instructional settings, including whole-class discussions, small-group problem-solving, and online lesson contexts (Suresh et al.,

[1] https://github.com/SumnerLab/TalkMoves/tree/main

2022). Transcripts were human-transcribed, segmented by speaker, and annotated at the utterance level with six pedagogically-grounded talk moves. These moves are organized into three higher-level dimensions: *Learning Community*, *Content Knowledge*, and *Rigorous Thinking* and include: (1) *Keeping Everyone Together*, *Getting Students to Relate to Another's Ideas*, and *Restating*; (2) *Pressing for Accuracy*; and (3) *Revoicing* and *Pressing for Reasoning*. The original dataset includes talk-move annotations for both teacher and student utterances; in this study, we focus only on teacher talk moves. We use a subset of 63 sessions, and provide detailed move definitions in Appendix B.

3.2 CLASS Annotation of Chat-based Math Tutoring Dialogue

The math tutoring dialogue dataset was collected and de-identified by an online platform that connects volunteer tutors with students attending predominantly low-income (Title I) schools in the United States.[2] Students initiate sessions on demand rather than following fixed schedules, and tutoring occurs through a synchronous text-based chat interface with an accompanying digital whiteboard. We focus on secondary mathematics tutoring. We randomly sampled 30 tutoring sessions for CLASS annotation by an expert who had completed formal training for applying the CLASS Instructional Support framework (Pianta et al., 2008, 2012). While CLASS was developed to evaluate how instructors support students' conceptual understanding and reasoning in classroom settings, we worked with the trained expert to adapt and operationalize the following subdimensions applicable to a one-to-one tutoring context: (1) *Feedback Loops*, (2) *Scaffolding* (3) *Building on Student Responses* and (4) *Encouragement and Affirmation*. Detailed definitions of these four moves appear in Appendix B. For brevity, we refer to this dataset as CLASS-annotated throughout the paper.

3.3 Re-Annotating Utterances with LLMs

For both datasets, the original utterance-level DA labels were produced by expert human annotators. Following recent protocols for LLM-assisted annotation workflows (Ahtisham et al., 2025), we additionally annotate the same transcripts with GPT-5 using LiteLLM API to obtain an LLM-generated label for each utterance. This yields two parallel rater-specific label sets per dataset, which lets us evaluate segmentation not only in terms of coherence-based criteria but also in terms of segment-level rater agreement: after aggregating utterance labels within each proposed segment into a segment-level label distribution, we measure how closely the human- and AI-derived segment distributions align under different segmentations. Full utterance-level prompts are provided in Appendix C.

[2]Platform name blinded for review.

4 Segmentation Experiment Setup

This section describes our task formulation and the four segmentation methods we evaluate. Two methods return boundaries using zero-shot prompting of LLMs, while the other two return boundaries using coherence-based objectives, including the Dial-Start algorithm (Gao et al., 2023). In each family, we compare a text-only variant and a DA-aware variant. For segmentation, all inputs consist only of the ordered dialogue text (utterance content) without speaker identifiers or any DA labels; DA annotations are used only for the Dial-Start DA-aware variant (as retrieval signals) and for evaluation. We chose these four methods to compare zero-shot LLM boundary proposal against a coherence-based baseline (Dial-Start), and to test whether adding DA information helps within each paradigm via matched text-only vs. DA-aware variants.

4.1 Task Definition

Given a dialogue as a sequence of utterances $\mathbf{u} = (u_1, \ldots, u_T)$, we predict a set of boundary indices $\mathcal{B} \subset \{1, \ldots, T-1\}$, where $j \in \mathcal{B}$ denotes a boundary placed *after* utterance u_j (i.e., between u_j and u_{j+1}). The boundaries induce $K = |\mathcal{B}|+1$ contiguous segments $S_1, \ldots, S_K$ such that each segment corresponds to a span in which the tutor's pedagogical intent (i.e., dialogue act definition) is relatively stable, and boundaries align with meaningful shifts in tutoring strategy.

4.2 Segmentation Methods

We study dialogue segmentation as an intermediate step before downstream annotation, comparing two method families: (i) LLM-based segmentation and (ii) unsupervised coherence-based segmentation. For each family, we evaluate a text-only variant and a DA-aware variant that incorporates the taxonomy, enabling controlled comparisons of how explicitly conditioning on pedagogical function affects boundary placement. We hypothesize that segmentation with awareness of the annotation goal as encoded in the taxonomy (a representation of the codebook) improves the quality of subsequent annotations.

4.2.1 LLM Segmentation (generic prompting)

A general-purpose LLM reads the full dialogue and outputs a set of boundary indices by marking perceived discourse or topic shifts. We ask it to segment the dialogue by outputting a JSON list of boundary indices (0-indexed turn numbers marking the last utterance in each segment). We use GPT-5 (OpenAI, 2025) and Gemini3-pro (Google, 2025) with fixed prompts and decoding settings. Prompts are provided in Appendix A.

4.2.2 LLM Segmentation (DA codebook prompting)

To incorporate the DA information, we prompt the same LLM with the codebook (i.e., DA definitions) and instruct it to place boundaries when the pedagogical function changes, using the codebook as explicit criteria for boundary decisions. The model is prompted to refer to the definitions of DAs and add a boundary when a shift occurs. All other settings are identical to the generic prompting condition. Prompt details are in Appendix A and DA definitions are in Appendix B.

4.2.3 Unsupervised Coherence Segmentation (Dial-Start)

As a non-LLM baseline, we use Dial-Start (Gao et al., 2023), an unsupervised dialogue segmenter that proposes boundaries from local coherence drops between neighboring utterances. Dial-Start trains an utterance encoder $f(\cdot)$ using a contrastive neighboring-utterance objective that scores adjacent pairs higher than mismatched pairs. At inference time, it computes a boundary score from changes in adjacent similarity and predicts $\mathcal{B}$ using the decoding protocol in Gao et al. (2023) and specific hyperparameters are described in Appendix D.

4.2.4 Unsupervised Coherence Segmentation with DA-conditioned Retrieval

To create a DA-aware variant of the coherence baseline (Dial-Start; Gao et al., 2023), we add a lightweight retrieval step inspired by retrieval-augmented methods (Lewis et al., 2020). Because our setting provides no gold segment boundaries, we do not train on boundary labels. Instead, we maintain a memory of expert-annotated human *DA-labeles* from the same dataset (1.9k labels for TalkMoves, 301 for CLASS-annotated) to inject semantic information about DAs. For each utterance

Symbol	Meaning
T	number of utterances in a dialogue
K	number of segments
S_k	k-th segment (contiguous utterances)
$\lvert S_k \rvert$	segment length
r	rater: Human (H) or AI (A)
C	number of DA categories
$y_i^{(r)}$	DA label for utterance u_i by rater r
$p_{k,c}^{(r)}$	proportion of DA c in segment k (rater r)
$\mathbf{p}_k^{(r)}$	DA distribution vector for segment k (rater r)
w_k	segment weight $\lvert S_k \rvert / T$
$\tilde{w}_k$	adjacent-pair weight $(\lvert S_k \rvert + \lvert S_{k+1} \rvert)/(2T)$

Table 2: Notation used in metric definitions.

u_i, we retrieve the top-K_{ret} semantically similar labeled utterances (by cosine similarity), aggregate their DA embeddings into a single DA vector r_i (via similarity-weighted averaging), and fuse it into the utterance representation:

$$\hat{h}_i = \text{norm}(h_i + \alpha r_i) \quad (1)$$

α is a fusion weight that balances the contribution of the retrieved DA embedding r_i against the original utterance representation h_i (see Appendix E for specific values used during inference).

We then replace h_i with $\hat{h}_i$ in Dial-Start's coherence computations. This injects codebook-aligned pedagogical signals into boundary prediction while leaving the underlying coherence objective unchanged. Details of DA-labeled memory and retrieval are provided in Appendix E.

Compute Resources. Dial-Start and its variant were run a workstation with two NVIDIA Quadro RTX 6000 GPUs (24 GB VRAM each). We used a single GPU and the models were implemented in PyTorch 2.9.1 (CUDA 12.8).

4.3 Evaluation

Most segmentation evaluation methods require gold labels for segments: P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002) are two widely used metrics for evaluating text and dialogue segmentation because they quantify boundary errors in a principled, sliding-window manner (Purver, 2011). However, gold-label segments are time-intensive to produce and suffer from some of the same ambiguity issues that affect utterance-level segmentation.

For this study, we propose a *gold-label-free* evaluation approach that does not depend on reference boundaries, and instead uses downstream, distributional indicators of segment coherence and separation to compare methods. To do so, we evaluate each segmentation by converting each segment into a distribution over DA labels. We focus on the *annotated DA label distributions* and measure (1) within-segment consistency, (2) between-segment distinctiveness, and (3) distribution differences between raters.

We first calculate the label distributions in each segment. A dialogue with T utterances is split into K contiguous segments $S_1, \ldots, S_K$. Let $r \in \{H, A\}$ denote the rater (Human vs. AI), and let there be C DA types. For segment S_k, we compute the empirical DA distribution

$$p_{k,c}^{(r)} = \frac{1}{|S_k|} \sum_{u_i \in S_k} I[y_i^{(r)} = c], \qquad \sum_{c=1}^{C} p_{k,c}^{(r)} = 1. \quad (2)$$

We weight segments by length: $w_k = |S_k|/T$. Notation is summarized in Table 2.

4.3.1 Consistency within a segment

Entropy (lower is better). Segments should have concentrated DA label distributions if they reflect a stable annotation pattern which, for our data, may represent a latent instructional method. We report a normalized entropy so that values lie in $[0, 1]$ by dividing by the maximum entropy $\log_2 C$.

$$\begin{aligned} H_k^{(r)} &= -\sum_{c=1}^{C} p_{k,c}^{(r)} \log_2 p_{k,c}^{(r)}, \\ \widetilde{H}_k^{(r)} &= \frac{H_k^{(r)}}{\log_2 C}, \\ \overline{\widetilde{H}}^{(r)} &= \sum_{k=1}^{K} w_k \, \widetilde{H}_k^{(r)}. \end{aligned} \quad (3)$$

Purity (higher is better). Purity is the share of the dominant DA label in the segment.

$$\mathrm{Pur}_k^{(r)} = \max_c p_{k,c}^{(r)}, \qquad \overline{\mathrm{Pur}}^{(r)} = \sum_{k=1}^{K} w_k \, \mathrm{Pur}_k^{(r)}. \quad (4)$$

4.3.2 Distinctiveness between adjacent segments

Good boundaries should separate segments with meaningfully different DA profiles. We measure distinctiveness in two complementary ways: a distributional shift between adjacent segments and a local label-transition rate at boundary positions.

Adjacent distribution shift via Jensen–Shannon divergence (higher is better). We compute the Jensen–Shannon (JS) divergence between DA label distributions of adjacent segments. For two distributions $\mathbf{p}$ and $\mathbf{q}$, JS is defined via the Kullback-Leibler divergence, $\mathrm{KL}(\mathbf{p}\|\mathbf{q}) = \sum_i p_i \log(p_i/q_i)$. Let $\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q})$; then:

$$\mathrm{JS}(\mathbf{p}, \mathbf{q}) = \tfrac{1}{2}\mathrm{KL}(\mathbf{p}\|\mathbf{m}) + \tfrac{1}{2}\mathrm{KL}(\mathbf{q}\|\mathbf{m}). \quad (5)$$

We aggregate adjacent distinctiveness under rater r as:

$$\overline{\mathrm{JS}}_{\mathrm{adj}}^{(r)} = \sum_{k=1}^{K-1} \tilde{w}_k \, \mathrm{JS}\Big(\mathbf{p}_k^{(r)}, \mathbf{p}_{k+1}^{(r)}\Big), \quad (6)$$

where $\tilde{w}_k = (|S_k| + |S_{k+1}|)/2T$ weights each boundary by the proportion of the dialogue it separates.

Boundary change rate (higher is better). As a simpler boundary-level diagnostic, we compute the fraction of predicted boundaries that coincide with a change in the utterance-level DA label:

$$\mathrm{BCR}^{(r)} = \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} I\Big[y_j^{(r)} \neq y_{j+1}^{(r)}\Big]. \quad (7)$$

4.3.3 Distributional divergence within a segment (lower is better)

We complement IRR with a distributional agreement metric between human and AI annotations by calculating JS divergence. If a segmentation yields coherent spans, then human and AI annotations should induce similar DA distributions within the *same* segments. We therefore compute

$$\overline{\mathrm{JS}}_{\mathrm{HA}} = \sum_{k=1}^{K} w_k \, \mathrm{JS}\Big(\mathbf{p}_k^{(H)}, \mathbf{p}_k^{(A)}\Big) \quad (8)$$

where lower values indicate closer agreement in segment-level label distributions.

5 Results

Table 3 reports average segment counts and gold-label-free quality metrics aligned with three goals: segment coherence (lower $\overline{\widetilde{H}}$ / higher $\overline{\mathrm{Pur}}$), boundary separation (higher $\overline{\mathrm{JS}}_{\mathrm{adj}}$ and BCR), and rater alignment (lower $\overline{\mathrm{JS}}_{\mathrm{HA}}$). Qualitative examples are provided in Appendix 8 and 9.

Two consistent patterns emerge. First, DA-awareness systematically improves consistency and distinctiveness for LLM-based segmenters but not for the coherence-based baseline. Second, LLM-based and Dial-Start segmenters optimize different and often competing objectives, with no single method dominating across metrics or datasets.

5.1 Segmentation granularity differs modestly across methods

Across both datasets, all methods produce a comparable number of segments per dialogue. These differences in average granularity (K) are modest relative to the high within-method variance observed in the results (Table 3). Although DA-aware prompting leads to a moderate increase in the number of segments for LLM-based models, particularly in

Table 3: Segmentation results using DA label distribution metrics, reported as mean [95% CI]. Means are rounded to 3 decimals and CI bounds to 2 decimals. Lower $\overline{H}$ and $\overline{\mathrm{JS}}_{\mathrm{HA}}$ indicate better within-segment consistency and human-AI agreement; higher $\overline{\mathrm{JS}}_{\mathrm{adj}}$ and BCR indicate stronger boundary distinctiveness. Granularity is reported as the average number of segments per dialogue (K; mean (SD))

	Granularity	**Consistency**		**Distinctiveness**		**Rater Agreement**
Method	K Mean (SD)	$\overline{\widetilde{H}}\downarrow$	$\overline{\mathrm{Pur}}\uparrow$	$\overline{\mathrm{JS}}_{\mathrm{adj}}\uparrow$	$\mathrm{BCR}\uparrow$	$\overline{\mathrm{JS}}_{\mathrm{HA}}\downarrow$
CLASS-annotated						
GPT-5	4.90 (1.71)	0.349 [0.24, 0.46]	0.546 [0.42, 0.67]	0.447 [0.31, 0.59]	0.222 [0.14, 0.31]	0.424 [0.31, 0.54]
GPT-5 DA-aware	6.30 (1.97)	**0.286** [0.18, 0.39]	**0.570** [0.44, 0.70]	0.477 [0.34, 0.62]	**0.288** [0.19, 0.38]	0.449 [0.33, 0.57]
Gemini-3-pro	4.47 (2.01)	0.384 [0.27, 0.50]	0.528 [0.41, 0.65]	0.447 [0.31, 0.59]	0.237 [0.13, 0.34]	**0.407** [0.30, 0.52]
Gemini-3-pro DA-aware	4.53 (1.70)	0.391 [0.27, 0.51]	0.531 [0.41, 0.65]	0.435 [0.29, 0.58]	0.267 [0.16, 0.38]	0.411 [0.30, 0.52]
Dial-Start	4.60 (0.56)	0.303 [0.19, 0.42]	0.564 [0.43, 0.70]	**0.545** [0.40, 0.69]	0.208 [0.12, 0.29]	0.459 [0.34, 0.58]
Dial-Start + DA-aware	4.50 (0.63)	0.319 [0.21, 0.43]	0.561 [0.43, 0.69]	0.515 [0.37, 0.66]	0.253 [0.15, 0.36]	0.484 [0.37, 0.60]
TalkMoves						
GPT-5	10.86 (4.87)	0.616 [0.55, 0.68]	0.659 [0.62, 0.70]	0.447 [0.37, 0.52]	0.235 [0.16, 0.31]	**0.470** [0.44, 0.50]
GPT-5 DA-aware	12.54 (6.97)	0.609 [0.55, 0.67]	0.664 [0.63, 0.70]	0.470 [0.41, 0.53]	0.222 [0.15, 0.29]	0.489 [0.46, 0.52]
Gemini-3-pro	16.62 (9.09)	0.598 [0.53, 0.66]	0.662 [0.63, 0.70]	0.471 [0.40, 0.54]	0.235 [0.16, 0.31]	0.480 [0.45, 0.52]
Gemini-3-pro DA-aware	19.53 (10.43)	**0.566** [0.50, 0.64]	**0.676** [0.64, 0.71]	**0.478** [0.41, 0.54]	0.252 [0.18, 0.33]	0.505 [0.47, 0.54]
Dial-Start	14.60 (7.27)	0.619 [0.56, 0.68]	0.640 [0.61, 0.67]	0.475 [0.42, 0.53]	0.416 [0.32, 0.52]	0.513 [0.48, 0.55]
Dial-Start + DA-aware	14.75 (7.18)	0.639 [0.58, 0.70]	0.633 [0.60, 0.66]	0.469 [0.42, 0.52]	**0.524** [0.46, 0.59]	0.503 [0.47, 0.54]

the TalkMoves dataset, it does not necessarily result in over-segmentation. This pattern suggests that the improvements in performance metrics are not a simple artifact of changes in granularity but instead reflect more precise boundary placement.

5.2 DA-awareness improves coherence for LLMs segmenters, but not for Dial-Start

Across datasets, incorporating DA information most consistently improves within-segment coherence for LLM-based segmenters. On CLASS-annotated, GPT-5 DA-aware yields the lowest normalized entropy and highest purity, outperforming its text-only counterpart and all non-LLM baselines. On TalkMoves the strongest coherence is achieved by Gemini-3-pro DA-aware, again improving over the text-only LLM variant (Table 3).

In contrast, DA-aware retrieval does not improve coherence for the coherence-based baseline. Dial-Start + DA-aware slightly degrades both entropy and purity relative to Dial-Start on both datasets, indicating that injecting DA information into a coherence objective does not make segments internally more homogeneous. Thus, DA-awareness acts as a coherence booster for LLM prompting, but not for coherence-based segmentation.

5.3 Boundary distinctiveness depends on the method family and dataset

Boundary quality exhibits a different pattern. On CLASS-annotated, Dial-Start achieves the strongest adjacent distinctiveness (highest $\overline{\mathrm{JS}}_{\mathrm{adj}}$), exceeding both LLM variants and Dial-Start + DA-aware. This suggests that coherence-based methods are particularly effective at separating segments with globally distinct DA distributions in this domain.

On TalkMoves, however, the strongest adjacent distinctiveness is achieved by Gemini-3-pro DA-aware, with Dial-Start close behind. Thus, when instructional moves are more varied and densely interleaved, DA-aware LLMs can rival or exceed coherence-based methods in separating distributionally distinct spans.

Local boundary contrast (BCR) further differentiates methods. On CLASS-annotated, GPT-5 DA-aware produces the highest BCR, indicating that its boundaries align most closely with local DA transitions. On TalkMoves, the highest BCR comes from Dial-Start + DA-aware, suggesting that retrieval is most effective for sharpening local, turn-level transitions in that dataset. Together, these results show that adjacent distinctiveness and local boundary contrast reward different segmentation behaviors and favor different method families.

5.4 No single segmenter dominates: a three-way trade-off

Across both datasets, no method simultaneously optimizes coherence, boundary separation, and human-AI distributional agreement. Improvements in coherence do not reliably translate into stronger boundary separation, and gains in either often coincide with worse human-AI alignment.

Notably, DA-aware prompting tends to increase human-AI JS divergence for LLMs, suggesting that codebook-guided boundary placement can reduce alignment with human-labeled distributions even as it improves internal coherence. In contrast, methods with stronger boundary distinctiveness often exhibit weaker rater alignment.

Overall, DA-awareness improves within-segment coherence for LLM-based segmenters, but does not consistently improve boundary quality or human–AI alignment. Across datasets, LLM-based methods tend to produce more construct-consistent spans, while coherence-based baselines more often yield sharper global boundaries.

6 Discussion and Conclusion

This study set out to answer two questions: whether incorporating dialogue-act (DA) information improves segmentation, and whether LLM-based segmentation is a viable alternative to coherence-based methods. DA-awareness consistently improves within-segment coherence for LLM-based segmenters, while LLM segmentation performs competitively with established coherence-based baselines. These gains are not uniform across evaluation criteria, suggesting segmentation is a multi-objective design problem rather than a single-score optimization task.

We find that improvements in within-segment coherence can conflict with boundary distinctiveness or alignment with human annotations. In pedagogical dialogue, this tension is amplified by the 'unitizing' problem, where annotators often agree on instructional intent but differ on its temporal extent (Beeferman et al., 1999; Pevzner and Hearst, 2002; Krippendorff et al., 2016; Mathet et al., 2015). As a result, segmentation methods should be selected with downstream goals in mind, such as identifying extended instructional phases versus detecting fine-grained shifts.

The benefits of DA-awareness are most evident in segment coherence. Across datasets, DA-aware prompting leads LLMs to produce more homogeneous, construct-consistent spans, often outperforming text-only LLMs and, in some settings, coherence-based baselines. This suggests that LLMs can internalize codebook-level definitions of pedagogical intent and use them to guide boundary placement even without gold segmentation labels. In contrast, injecting DA information into a coherence-based objective via retrieval does not yield similar gains, indicating DA-awareness interacts differently with instruction-following models than with similarity-based criteria.

Boundary quality depends strongly on both the segmentation paradigm and the dataset. Coherence-based methods tend to place boundaries that separate globally distinct regions of dialogue, whereas LLM-based methods, especially when DA-aware, are better at carving internally consistent spans aligned with instructional constructs. Different notions of boundary success also favor different methods, and this divergence underscores that good segmentation is task-dependent.

A central empirical finding is that improvements in coherence often coincide with reduced alignment to human-labeled distributions. DA-aware prompting frequently increases human–AI divergence, revealing a tension between enforcing a formal codebook and reproducing human annotation patterns. Human annotators often rely on pragmatic judgment and contextual smoothing, implicitly tolerating boundary ambiguity in order to preserve conversational flow. By contrast, a codebook-injected LLM applies the taxonomy more literally, producing boundaries that are theoretically consistent but less sensitive to interactional nuance.

From this perspective, divergence between human and LLM segmentations should not be interpreted solely as model error. Instead, it highlights segmentation as an explicit modeling decision that shapes how constructs are operationalized. LLMs, when guided by a fixed codebook, can serve as diagnostic tools: by enforcing a stable interpretation of criteria, they can surface latent ambiguities, theoretical drift, or underspecified boundary conventions in human-annotated datasets.

Overall, our results show that segmentation should be treated as a first-class design choice and reported with the downstream use case in mind. DA-aware LLM segmentation is well suited for producing coherent spans that closely follow a formal construct definition, while coherence-based methods remain strong for detecting sharp shifts in dialogue flow. Because these approaches optimize different aspects of segmentation quality, we recommend evaluating and selecting segmenters using criteria that match the intended analysis, rather than assuming a single best method.

7 Limitations

Our evaluation intentionally avoids gold segment boundaries, instead using distributional criteria to assess segmentation quality without reference labels. These metrics may not capture all pedagogically meaningful shifts, particularly those requiring domain expertise. The results also depend on the underlying DA taxonomies and labeling quality, and systematic differences between human and LLM annotations can influence agreement-based measures. Esepeically for the CLASS-annotated dataset, which contains very few labeled utterances. This low density may limit the resolution of our metrics and disproportionately influence human-AI agreement measures compared to more densely labeled corpora. Finally, the study is limited to two datasets and a restricted set of prompting and retrieval designs, and future work is needed to assess generalization to multi-party or multimodal settings.

8 Ethical Considerations

We analyze de-identified tutoring dialogue datasets in accordance with our Institutional Review Board (IRB)-approved protocol [3]. TalkMoves dataset was de-identified by humans before it was open sourced and CLASS-annotated dataset was de-identified by humans by the tutoring provider and the participants consented to its use for research. We follow IRB-approved procedures for data storage and access. We use these data solely for research purposes to understand and improve dialogue annotation workflows and analyze instructional interactions. Re-identification attempts, user profiling, or any use that could enable harm to individual participants is unacceptable. Although the data are de-identified, we recognize that conversational text can still carry residual privacy risk; we therefore minimize the inclusion of verbatim excerpts, avoid reporting sensitive attributes, and present results in aggregate wherever possible. Any inferences drawn from these datasets should be interpreted cautiously, as they reflect the context and population captured by the underlying platforms and may not generalize to other settings.

References

Bakhtawar Ahtisham, Kirk Vanacore, Jinsook Lee, Zhuqian Zhou, Doug Pietrzak, and Rene F. Kizilcec. 2025. Ai annotation orchestration: Evaluating llm verifiers to improve the quality of llm annotations in learning analytics. *Preprint*, arXiv:2511.09785.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

[3] IRB number blinded for review.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, pages 26–33.

Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. 2024. S3-dst: Structured open-domain dialogue segmentation and state tracking in the era of llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14996–15014.

Nathan Duran, Steve Battle, and Jim Smith. 2022. Inter-annotator agreement using the conversation analysis modelling schema, for dialogue. *Communication Methods and Measures*, 16(3):182–214.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, pages 334–343.

Chong Fan, Brian Mak, and Stephen Wan. 2024. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*, pages 562–569.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware contrastive learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2481–2485.

Google. 2025. Gemini 3 Pro. `https://gemini.google.com/`. Version used: [Specific version date or name, e.g., March 14 version]; Accessed: [Date accessed].

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, culture and social interaction*, 9:16–44.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. Superdialseg: A large-scale dataset for supervised dialogue segmentation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. Meganno+: A human-llm collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176.

Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of NAACL*, pages 469–473.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. In *International Conference of the Italian Association for Artificial Intelligence*, pages 151–171. Springer.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL*, pages 25–32.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

T Daniel Midgley. 2009. Dialogue segmentation with large numbers of volunteer internet annotators. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 897–904.

Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappé. 2011. Text segmentation via topic modeling: an analytical study. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1553–1556.

Hadi Mohammadi, Tina Shahedi, Pablo Mosteiro, Massimo Poesio, Ayoub Bagheri, and Anastasia Giachanou. 2025. Assessing the reliability of LLMs annotations in the context of demographic bias and model explanation. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 92–104, Vienna, Austria. Association for Computational Linguistics.

Ying Na and Shihui Feng. 2025. Llm-assisted automated deductive coding of dialogue data: leveraging dialogue-specific characteristics to enhance contextual understanding. In *International Conference on Artificial Intelligence in Education*, pages 248–262. Springer.

OpenAI. 2025. ChatGPT-5 Large language model. `https://openai.com/gpt-5/`. Accessed: 22 December 2025.

Lidiia Ostyakova, M Molchanova, Ksenia Petukhova, Nika Smilga, Danila Kornev, and Mikhail Burtsev. 2022. Corpus with speech function annotation: Challenges, advantages, and limitations. *Computational Linguistics and Intellectual Technologies*, pages 1129–1139.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. Chatgpt vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254.

Kseniia Petukhova and Ekaterina Kochmar. 2025. Intent matters: Enhancing ai tutoring with fine-grained pedagogical intent annotation. *arXiv preprint arXiv:2506.07626*.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. 2008. *Classroom Assessment Scoring System (CLASS) Manual, K–3*. Paul H. Brookes Publishing, Baltimore, MD.

Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. 2012. *Dimensions Guide*. Teachstone, Charlottesville, VA. Practical guide to the CLASS framework.

Matthew Purver. 2011. Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, pages 291–317.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs understand dialogues? a case study on dialogue acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria. Association for Computational Linguistics.

Changyong Qi, Longwei Zheng, Yuang Wei, Haoxin Xu, Peiji Chen, and Xiaoqing Gu. 2024. Edudcm: A novel framework for automatic educational dialogue classification dataset construction via distant supervision and large language models. *Applied Sciences*, 15(1):154.

Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, and 1 others. 2025. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Transactions on the Web*, 19(3):1–29.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Hang Su and Jun Ye. 2025. Large language models for automating fine-grained speech act annotation: A critical evaluation of gpt-4o and deepseek. *Corpus Pragmatics*, pages 1–20.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. *arXiv preprint arXiv:2204.09652*.

Leila Tavakoli and Hamed Zamani. 2025. Reliable annotations with less effort: Evaluating llm-human collaboration in search clarifications. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 92–102, New York, NY, USA. Association for Computing Machinery.

Darinka Verdonik. 2023. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28(2):144–171.

Rose E Wang, Pawan Wirawarn, Kenny Lam, Omar Khattab, and Dorottya Demszky. 2024. Problem-oriented segmentation and retrieval: Case study on tutoring conversations. *arXiv preprint arXiv:2411.07598*.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of ACL*, pages 626–636.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, 29(4):534–561.

A LLM Segmentation Prompts

Table 4 summarizes the prompts used for LLM-based segmentation, comparing a zero-shot topic segmentation setting with a DA-aware setting. For the DA-aware condition, the prompt included DA (i.e. tutoring move) definitions from the codebook to guide boundary decisions. Details of definitions are described in Appendix B.

B Dialogue Act (DA) definitions

B.1 CLASS-annotated dataset

Subcategory 1: Feedback Loops Focus on the quality of feedback in the tutoring session, which is defined as the degree to which feedback advances learning and understanding and encourages student participation. A feedback loop refers to a sustained back-and-forth exchange between a teacher and a student that builds on prior turns and pushes learning forward.

- Low Quality: Feedback is mostly absent, minimal, or one-sided. Exchanges are brief, superficial, and do not build on student thinking.
- Mid Quality: Feedback loops occur occasionally, but are inconsistent or limited in depth. Some back-and-forth exchanges involve clarifying questions or brief elaboration.
- High Quality: Feedback loops are frequent and often extend or deepen student understanding. The teacher regularly engages in meaningful back-and-forth exchanges that build on student thinking.

Subcategory 2: Scaffolding Scaffolding refers to the teacher's use of hints, prompts, or structured support that helps students move toward successful task completion or deeper understanding. It is not however an explanation of a concept. Scaffolding enables students to perform at a higher level than they could independently by breaking down tasks, modeling thinking, or guiding problem-solving steps.

- Low Quality: Students are not provided with meaningful assistance, hints, or prompting, and are often left to complete work on their own.
- Mid Quality: The teacher sometimes scaffolds student learning, but these interactions are often brief or shallow.
- High Quality: The teacher frequently scaffolds student learning, enabling students to perform at a higher level than they could independently. Support is purposeful and leads to visible progress.

Subcategory 3: Building on Student Responses Building on student responses refers to the teacher's practice of expanding, instances of clarification and specific feedback, or refining what a student says to further their understanding. This includes asking targeted follow-up questions, elaborating on partial answers, or redirecting thinking in a way that pushes learning forward. It is not praise, repetition, or scaffolding.

- Low Quality: The teacher gives no or very general feedback. They accept student answers and move on without clarifying or extending.
- Mid Quality: The teacher occasionally expands or clarifies student responses, but these exchanges are brief or shallow.
- High Quality: The teacher frequently builds on student responses with meaningful expansions. This includes giving specific feedback, asking clarifying or elaborative questions, and encouraging deeper reasoning.

Subcategory 4: Encouragement and Affirmation Encouragement and affirmation refer to the teacher's recognition of student effort and encouragement of persistence, often through specific praise or motivational feedback that supports continued engagement in the learning process. Generic statements such as "Great," "Okay," "That's correct," or "Glad I could help" do not count as encouragement or affirmation.

- Low Quality: The teacher gives little or no encouragement. Feedback is correctness-focused and does not recognize student effort or struggle.
- Mid Quality: The teacher occasionally encourages students with brief, general comments (e.g., "Nice try," "Keep going").
- High Quality: Encouragement targets student effort, strategies, or perseverance. It helps students persist during challenges and feel supported.

Prompt 1: Zero-shot topic segmentation
You are an expert educational discourse analyst. Your task is to segment a tutoring dialogue into coherent topics. **Definition:** • Boundary index = the 0-indexed turn number of the last utterance in a segment. • Segments must be contiguous and cover the entire dialogue. • Always include the final turn index as the last boundary. **Add a boundary when:** • The dialogue moves to a new problem/question, or a new subtask with a different goal. • The interaction shifts between social talk and lesson content. **Do NOT add a boundary when:** • The tutor restates the same idea, checks understanding, or the student gives brief answers within the same topic. **Output format** Return ONLY a JSON object in the following form (no extra text, no markdown): `{"boundary_indices":[integer, integer, ...]}`
Prompt 2: DA-aware segmentation
You are an expert educational discourse analyst. Your task is to segment a tutoring dialogue into coherent topics. **Definition:** • Boundary index = the 0-indexed turn number of the last utterance in a segment. • Segments must be contiguous and cover the entire dialogue. • Always include the final turn index as the last boundary. **Add a boundary when:** • The dialogue moves to a new problem/question, or a new subtask with a different goal. • The interaction shifts between social talk and lesson content. **Do NOT add a boundary when:** • The tutor restates the same idea, checks understanding, or the student gives brief answers within the same topic. Use the following construct definitions to decide where meaningful boundaries occur. Do NOT label the dialogue with these constructs; use them only to guide segmentation. **[ADD Move Definitions]** **Output format** Return ONLY a JSON object in the following form (no extra text, no markdown): `{"boundary_indices":[integer, integer, ...]}`

Table 4: System prompts used for LLM-based segmentation: (1) zero-shot topic segmentation and (2) DA-aware segmentation.

B.2 TalkMoves dataset

Subcategory 1: Keeping everyone together Prompting students to be active listeners and orienting students to each other

Subcategory 2: Getting students to relate to another's ideas Prompting students to react to what a classmate said

Subcategory 3: Restating Repeating all or part of what a student said word for word

Subcategory 4: Pressing for accuracy Prompting students to make a mathematical contribution or use mathematical language

Subcategory 5: Revoicing Repeating what a student said but adding on or changing the wording

Subcategory 6: Pressing for reasoning Prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations

C Annotation prompt at the utterance level

Table 5 and 6 shows the prompt used to generate utterance-level labels with an LLM. We used GPT-5 to produce the annotation outputs.

Key	Value (verbatim)
Role	You are an expert educational discourse analyst. Your task is to label each teacher utterance (Speaker = T) with one Talk Move from the Open-Source Talk Moves Codebook. Use the coding definitions and examples from the manual to guide your decision. Provide a concise explanation in the 'Reasoning' field only when a Talk Move is present.
Workflow	Read the dialogue carefully. For each teacher utterance (Speaker = T), assign exactly ONE Talk Move from the Allowed Moves list. If an utterance could fit multiple moves, choose the one that best represents the communicative function in context. If no Talk Move applies (e.g., the utterance is off-topic, evaluative, unclear, or unrelated to the lesson), assign the label 'None' and leave the 'Reasoning' field empty/null. Do NOT generate reasoning for 'None' labels.
AllowedMoves	Keeping Everyone Together Getting Students to Relate to Another's Ideas Restating Pressing for Accuracy Revoicing Pressing for Reasoning
MoveDefinitions	"Keeping Everyone Together": "Teacher prompts students to be active listeners and orienting students to each other." "Getting Students to Relate to Another's Ideas": "Teacher prompts students to react to what a classmate said." "Restating": "Teacher repeats all or part of what a student said word for word." "Pressing for Accuracy": "Teacher prompts students to make a mathematical contribution or use mathematical language." "Revoicing": "Teacher Repeats what a student said but adding on or changing the wording." "Pressing for Reasoning": "Teacher prompts students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations."
MoveExamples	"Keeping Everyone Together": ["So x equals five dollars, right?", "It's going to be 150, right?", "Are you finished?"] "Getting Students to Relate to Another's Ideas": ["How do you feel about what they said?", "Does anyone understand how she solved the problem?", "Do you agree or disagree with him?"] "Restating": ["S: The same size and shape but it moves to a different position. \n T: It moves to a different position. (Restating)", "S: An exponent \n T: Exponent. (Restating)", "S: It's four million and then the two. \n T: Four million two. (Restating)"] "Pressing for Accuracy": ["What is the answer to number 2?", "How did you solve it?", "What does x stand for?"] "Revoicing": [" S: It had two. \n T: So instead of one flat edge, it had two. (Revoicing)", " S: Oh, Company B. \n T: It's Company B because that's the one that charges you $2.00 per minute. (Revoicing)", " S: I got 2X minus Y. \n T: It's not 2X. (Revoicing)", " S: La respuesta es siete. \n T: The answer is seven. (Revoicing)"] "Pressing for Reasoning": ["Can you explain why?", "How are these ideas connected?", "Where do we see the x + 1 in the tiles?"]
RequiredOutput	format: JSON envelope: Respond with a single JSON object with a top-level field named 'records'. fields: ID (integer, exactly copied from the input utterance - THIS IS REQUIRED for matching); Speaker (string, exactly copied from the input utterance (usually 'T')); Turn (integer, exactly copied from the input utterance); TalkMove (one of the six Allowed Moves as a string, or null if no move applies.); Reasoning (string explanation for the chosen TalkMove, or null if no move applies.) example: {"records":[{"ID":1902,"Speaker":"T","Turn":150,"TalkMove":"Pressing for Reasoning","Reasoning":"The teacher asks students to explain whether a representation is correct, prompting justification."},{"ID":1905,"Speaker":"T","Turn":153,"TalkMove":null,"Reasoning":null}]}

Table 5: System prompt for utterance-level annotation (TalkMoves)

Key	Value (verbatim)
Role	You are an expert in developmental and educational psychology, with advanced training using classroom observation protocols such as the Classroom Assessment Scoring System (CLASS). You will read, analyze, and annotate transcripts of teacher-student tutoring conversations for evidence of effective interactions that promote student learning.
CategoryPrompt	Focus on the quality of feedback in the tutoring session, which is defined as the degree to which feedback advances learning and understanding and encourages student participation.
Subcategory1_Name	Feedback Loops
Subcategory1_Prompt	Your task is to identify individual utterances during the tutoring session that demonstrate high-quality feedback loops. Return a value of 1 at the start of a feedback loop, leave it blank otherwise.
Subcategory1_Definition	A feedback loop refers to a sustained back-and-forth exchange between a teacher and a student that builds on prior turns and pushes learning forward. Identify feedback loops that enable the student to achieve a deeper understanding of the material and concepts. In these interactions, the teacher should build upon the student's initial comment, engaging the student in sustained discussion aimed at deepening comprehension of the content.
Subcategory1_Examples	1. Teacher: "What do you need to do first?" Student: "Find the diameter." Teacher: "Why start there?" Student: "It's part of the formula." Teacher: "Right—so how will you use it?" 2. The teacher tells a student to include units in her answer for a math problem. Student: "Why do we have to write the units?" Teacher: "Who's going to read your answer?" Student: "You are." Teacher: "Look at the instructions. It says your answer should be clear to anyone, even someone who hasn't seen the problem." Student: "But everyone knows it's talking about centimeters." Teacher: "It might feel that way because everyone in our class is working on the same problem, but someone else reading your answer might not know what you're measuring, and the units help them understand." Student: "I guess writing the units would help someone who hasn't done the problem." 3. When a student gives an incorrect answer to a problem, the teacher keeps asking follow-up questions to help the student work through their thinking until they show they understand the concept. 4. A student mentions that making calculation errors is a problem in math class. The teacher asks, "What kinds of mistakes are people making in their calculations?" The student answers, "Adding instead of subtracting, or forgetting to carry numbers." The teacher asks, "Why do you think they're making these mistakes?"
Subcategory2_Name	Scaffolding
Subcategory2_Prompt	Your task is to identify individual utterances during the tutoring session that demonstrate high-quality Scaffolding. Return a value of 1 at the start of a Scaffolding interaction; leave it blank otherwise.
Subcategory2_Definition	Scaffolding refers to the teacher's use of hints, prompts, or structured support that helps students move toward successful task completion or deeper understanding. It enables students to perform at a higher level than they could independently by breaking down tasks, modeling thinking, or guiding problem-solving steps.
Subcategory2_Examples	1. The teacher notices a student who is struggling and says to her, "Let's go back and look at the relationship between the diameter and circumference of a circle." 2. "If you are having difficulty with this problem, remember that it is just like the ones we did earlier." 3. Teacher: "What do you need to do first? Next?" The student gives an answer. Teacher asks, "Can you explain why you think that?" or "How did you get that answer?"
Subcategory3_Name	Building on student responses
Subcategory3_Prompt	Your task is to identify individual utterances during the tutoring session that demonstrate high-quality building on student responses. Return a value of 1 at the start of a building on student response interaction; leave it blank otherwise.
Subcategory3_Definition	Building on student responses refers to the teacher's practice of expanding, clarifying, or refining what a student says to further their understanding. This includes asking targeted follow-up questions, elaborating on partial answers, or redirecting thinking in a way that pushes learning forward.
Subcategory3_Examples	1. "You're right, that's a pattern. And in this case, it repeats every three steps, which means it's a repeating pattern." 2. When a student says, "The answer is 16 because I multiplied 4 and 4," the teacher responds, "Multiplying 4 by 4 does give you 16—that's one way to think about it. But you could also get 16 by adding 8 and 8. What do you notice about both of those numbers?" 3. When a student says, "A system is just when everything is random," the teacher replies, "Not exactly. Remember, a system is a set of parts that work together in an organized way. Can you think of something in everyday life that works like that?"
Subcategory4_Name	Encouragement and Affirmation
Subcategory4_Prompt	Your task is to identify individual utterances during the tutoring session that demonstrate high-quality affirmation and encouragement. Return a value of 1 at the start of a affirmation and encouragement interaction; leave it blank otherwise.
Subcategory4_Definition	Encouragement and affirmation refer to the teacher's recognition of student effort and encouragement of persistence, often through specific praise or motivational feedback that supports continued engagement in the learning process.
Subcategory4_Examples	1. "That was so good! That was an incredible explanation of how you solved that problem." 2. "Those are some really great ideas. You're really cranking them out!" 3. "Figure it out. Don't give up. You can do it!" 4. "This is a really tough topic you've picked. I am so impressed by how hard you are working. I can't wait to see the finished project!"
RequiredOutput	Return a value of 1 at the start of the target interaction; leave it blank otherwise.

Table 6: System prompt for utterance-level annotation (CLASS-annotated)

D Hyperparameters of Dial-Start and its variant

Table 7 summarizes the hyperparameters that control boundary selection in the coherence-based baseline. These parameters determine how boundary candidates are scored, how many boundaries can be selected per dialogue, and how close two boundaries are allowed to be.

E DA-labeled memory and retrieval

Let $\mathcal{D}_{\text{mem}}$ denote a pool of utterances with move labels $m \in \{1, \ldots, M\}$ (e.g., human-annotated

Parameter	Value	Role in boundary selection
`window_size`	2	Context window used when scoring each candidate boundary by comparing left/right dialogue context; larger values incorporate broader surrounding context.
`alpha`	0.5	Threshold scale for selecting boundary candidates from depth scores: $\texttt{thr} = \mu + \alpha\sigma$; higher α yields fewer boundaries (stricter cutoff).
`pick_num`	4	Maximum number of boundaries returned per dialogue when `avg_seg_len` is unset (hard cap on selected boundary indices).
`avg_seg_len`	None	If set, defines an implicit cap on boundaries via an estimated segment count $\hat{K} \approx \lceil T/\texttt{avg\_seg\_len} \rceil$ (controls average segment length).
`min_gap`	3	Minimum separation (in utterances) enforced between consecutive boundaries to avoid trivially short adjacent segments.

Table 7: Key hyperparameters for the coherence-based segmentation baseline's boundary selection.

TalkMoves/CLASS-annotated labels). We construct a memory bank of labeled embeddings

$$\mathcal{M} = \{(h_j, m_j)\}_{j=1}^{N_{\text{mem}}}, \quad h_j = f(u_j),$$

and, for each query utterance u_i with embedding h_i, retrieve the top-K_{ret} neighbors by cosine similarity:

$$\{(h_{j_k}, m_{j_k})\}_{k=1}^{K_{\text{ret}}} = \text{TopK}_{(h,m)\in\mathcal{M}} \ \cos(h_i, h).$$

We convert similarities into attention weights via a temperature-τ softmax,

$$a_k = \frac{\exp(\cos(h_i, h_{j_k})/\tau)}{\sum_{\ell=1}^{K_{\text{ret}}} \exp(\cos(h_i, h_{j_\ell})/\tau)},$$

and form an aggregated move vector using a learned move-embedding table $E \in R^{M\times d}$ with row vectors e_m:

$$r_i = \sum_{k=1}^{K_{\text{ret}}} a_k \, e_{m_{j_k}}, \qquad \hat{h}_i = \text{norm}(h_i + \alpha r_i).$$

We then replace h_i with $\hat{h}_i$ in Dial-Start's coherence computations for boundary scoring.

F Example of Segmentation Result

F.1 TalkMoves

Table 8 describes an example dialogue in TalkMoves dataset. The following indices mark the turn after which each model placed a boundary:

- **GPT-5**: 0, 23, 24, 60, 76
- **GPT-5 DA-aware**: 0, 12, 18, 23, 24, 39, 59, 76
- **Gemini Pro 3**: 0, 59, 76
- **Gemini Pro 3 DA-aware**: 0, 12, 18, 39, 58, 59, 76
- **Dial-Start**: 11, 44
- **Dial-Start + DA-aware**: 14, 44

F.2 CLASS-annotated

Table 9 describes an example dialogue in CLASS-annotated dataset provided by a math tutoring platform[4]. The following indices mark the turn after which each model placed a boundary:

- **GPT-5**: 0, 8, 23, 30, 37
- **GPT-5 DA-aware**: 8, 15, 23, 27, 31, 33, 37
- **Gemini Pro 3**: 2, 34, 37
- **Gemini Pro 3 DA-aware**: 8, 30, 37
- **Dial-Start**: 7, 14, 22, 31
- **Dial-Start + DA-aware**: 2, 7, 17

[4] Platform name blinded for review

Table 8: An Example of Segmentation with TalkMoves dataset

Spkr	Utterance Content	Human Label	GPT-5 Label	ID
T	T and I were just chatting a little bit that its interesting for us to see all of your thinking because there are multiple ways of thinking about these problems and there are multiple ways of solving and getting the same answer			0
T	I would like to know if I look at this problem the six and a half because I know that I can do six minus four right but I cant do one twelfth minus seven twelfths it wont work because I only have one twelfth I cant take away seven twelfths			1
T	Is there a way that I could change six twelfths to help me	Keeping Everyone Together	Pressing for Accuracy	2
T	Is there a way for me to change six twelfths	Keeping Everyone Together	Pressing for Accuracy	3
T	I mean six and one twelfths sorry			4
T	How did you get seventy three twelfths	Pressing for Accuracy	Pressing for Reasoning	5
S	Well twelfth twelfths equals one whole so I did six by twelve			6
T	Then what did you do	Pressing for Accuracy	Pressing for Reasoning	7
S	Then I			8
T	Four and ten and twelve	Keeping Everyone Together		9
T	You know what			10
T	You did is exactly what Gavin and T were taking about over here exactly		Keeping Everyone Together	11
T	He changed these two into all twelfths before he subtracted them and that would work		Revoicing	12
T	Is there a different thing that we could do	Keeping Everyone Together	Pressing for Reasoning	13
T	Alyssa	Keeping Everyone Together		14
S	I did six twelfths so I did twelve twelfths and twelve twelfths plus twelve twelfths plus twelve twelfths plus twelve twelfths plus twelve twelfths and then I did plus one twelfth minus twelve twelfths minus twelve twelfths minus twelve twelfths and then I did minus seven twelfths			15
T	She broke then all into twelfths		Revoicing	16
S	Then I at the top I minused the four and then I took off a fifth one and then I did seven plus what equals twelve which is five so I put it at five twelfths			17
S	Then I did five twelfths plus one twelfth to get one and sixth twelfths			18
T	Okay now can I show you something			19
T	What if I treated one of these wholes what if I took a whole away from this and made it twelve twelfths	Pressing for Accuracy	Pressing for Reasoning	20
S	What you could do is you could take off two wholes and make it four and then you could minus the four			21
T	Okay so I want you to look at this		Keeping Everyone Together	22
T	Is five plus twelve twelfths going to give me six	Keeping Everyone Together	Pressing for Accuracy	23
T	Lindon eyes up here buddy		Keeping Everyone Together	24
T	Five plus twelve twelfths is going to give me six right plus I have another twelfth here so how many twelfths is that	Pressing for Accuracy	Pressing for Accuracy	25
T	Right so now watch		Keeping Everyone Together	26
T	I have five and thirteen twelfths now can I subtract my fractions	Keeping Everyone Together	Pressing for Accuracy	27
T	Look at that		Keeping Everyone Together	28
T	Okay so now I have five and thirteen twelfths minus four and seven twelfths			29
T	I can subtract those now right	Keeping Everyone Together	Pressing for Accuracy	30
T	Okay so whats my answer	Pressing for Accuracy	Pressing for Accuracy	31
T	Five minus four is	Pressing for Accuracy	Pressing for Accuracy	32
T	One and thirteen minus seven is	Pressing for Accuracy	Pressing for Accuracy	33
T	I want you to look again at what I did because I couldnt do one minus seven I had to take this and I had to trade in one of my wholes		Keeping Everyone Together	34
T	I had to break my whole up just like I did with that one when I had the circles here			35
T	I broke it up into fourths so I could take some away			36
T	Its the same as this			37
T	I took one of my wholes and broke it into twelve pieces and then I figured out six and one twelfth is the same as five and thirteen twelfths			38
T	I had to turn it into an improper fraction and then I subtract it			39
T	T do you have anything you want to add to that		Getting Students to Relate to Another's Ideas	40
T	Yes because I think about it in just a little bit of a different way			41
T	Do you want to see how I think about it		Keeping Everyone Together	42
T	I think you think about it like Dylan did because you shared Gavin			43
T	Just like Alyssa did			44
T	Yes because if you have twelve twelfths six times thats the same as six wholes			45
T	Yes			46
T	No			47
T	We should try it on another one			48
T	Were getting it			49
T	Wait wait wait what			50
T	Guys still need three more			51
T	I need twelve twelfths six times			52
T	And twelve times six is seventy two			53
T	Plus another one twelfth yes			54
T	So you had seventy three twelfths right		Revoicing	55
T	Thats how I just did it with Gavin			56
T	This is seventy three twelfths written out broken down			57
T	Thats exactly what Alyssa did too			58
T	Guys do you see the way that I would think of it versus the way T would think of it the way that Alyssa Dylan Gavin my brain thinks differently and thats okay because we would all end up at the same answer right		Keeping Everyone Together	59
T	I think lets go to the next one			60
T	Yes			61
T	How do you feel about doing them pair up and they talk about it with			62
T	Sure so this was the last one on here but we can do it in their journals			63
T	Do you think theyre going to need their board to do the work			64
T	There is a spot in here			65
T	Let me make sure theres enough space			66
T	Yes theres some space in here so T do you want to write page 171 up there			67
T	Up here			68
T	Yes			69
T	Its all right you can just write it			70
T	You know what I can do I can pull up the journal page on here			71
T	Heres the thing you might or might not use your board you could erase it			72
T	Were going to have to do another follow up but its a journal follow up			73
T	Oh yes page 171			74
T	This is a journal problem and it looks like they have gave you space in your journal			75
T	If you need your whiteboard you are welcome to still use it			76

Table 9: An Example of Segmentation with a tutoring provider's dataset annotated with CLASS framework

Spkr	Utterance Content	Human Label (Move)	GPT-5 Label	ID
TEACHER	hi what can i help you with today?			0
STUDENT	hi i need help on my math homework			1
STUDENT	it's on the factor theorem and the remainder theorem.			2
STUDENT	it's as follows:			3
STUDENT	Create a polynomial $p(x)$ which has the desired characteristics. You may leave the polynomial in factored form.			4
STUDENT	(i) the zeros of p are $c = 2$, and $c = -2$, and $c = -1$, and $c = 1$			5
STUDENT	(ii) The leading term of $p(x)$ is $117x^2$			6
STUDENT	There			7
STUDENT	How do i do this?			8
TEACHER	okay			9
TEACHER	is this a single problem?			10
TEACHER	since there are 4 zeroes, the leading term should be x^4	Building on student responses	Scaffolding	11
STUDENT	yeah i am confused by that too			12
TEACHER	do you think it could be a typo		Scaffolding	13
STUDENT	i dont think so, my teacher didn't say anything about it			14
TEACHER	i think it is impossible to have 4 roots with a quadratic equation		Scaffolding	15
STUDENT	ill draw out what i tried			16
STUDENT	actually it was someone else's idea but neither she nor i have any idea if it's right			17
TEACHER	hm			18
TEACHER	the only problem i have with this is that it does not look like a polynomial function	Building on student responses	Scaffolding	19
STUDENT	oh okay. do you have any other ideas? im kind of lost			20
TEACHER	honestly i think it's just a typo			21
TEACHER	maybe she meant any type of function			22
TEACHER	because then this would be the answer		Scaffolding	23
STUDENT	what kind of function is this then?		Feeback Loop	24
TEACHER	a rational function		Feeback Loop	25
STUDENT	what is a rational function?		Feeback Loop	26
TEACHER	a rational function is a function with polynomials on the numerator and denominator		Scaffolding	27
STUDENT	so no other ideas?			28
STUDENT	if it isn't a typo i mean			29
TEACHER	i just don't think it's possible			30
STUDENT	oh okay then. thanks for trying.	Feeback Loop		31
TEACHER	sorry im not the teacher so i cant tell you if it should be something else			32
TEACHER	but if it doesnt fit the rules, dont be afraid to question the question	Encouragement and affirmation	Scaffolding	33
STUDENT	oh okay. i will keep that in mind.			34
STUDENT	im going to end the session now.			35
STUDENT	thanks!			36
TEACHER	okay come back if you have any other questions!	Encouragement and affirmation		37

G Responsible NLP Research Checklist

A1. Did you describe the limitations of your work? Yes. Please see Section 7.

A2. Did you discuss any potential risks of your work? Yes. Please see Section 8.

B. Did you use or create scientific artifacts? Yes. We used an open-source dataset (TalkMoves), a math tutoring dataset from an online platform (CLASS-annotated) and Dial-Start model.

B1. Did you cite the creators of artifacts you used? Yes. Please see Section 3 and 4.2.

B2. Did you discuss the license or terms for use and / or distribution of any artifacts?

- TalkMoves : CC BY-NC-SA 4.0
- CLASS-annotated : The tutoring provider agreed on using the dataset and the participants consented to its use for research.
- Dial-Start: MIT License `https://github.com/AlibabaResearch/DAMO-ConvAI/blob/main/LICENSE`

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)? Yes. Please see Section 3 and 8.

B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it? Yes. Please see Section 3 and 8.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? Yes. Please see Section 3.

B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Yes. Please see Section 3.

C. Did you run computational experiments? Yes.

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? Yes. Please see Section 4.2.4.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? Yes. Please see Section 4 and Appendix A - E.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? Yes. Please see Table 3

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? Yes, we used NLTK, sklearn, torch, transformers. Please see Section 4 and Appendix D.

D. Did you use human annotators (e.g., crowdworkers) or research with human participants? Yes. Please see Section 3.2.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? Yes. Please see Section 3.2, 8.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)? Yes. Please see Section 3.2.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. Please see Section 8.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board? Yes. Please see Section 8.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? Yes. Please see Section 3.2.

E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing? Yes.

E1. If you used any AI assistants, did you include information about your use? The authors used AI assistants (ChatGPT, Google Scholar Labs, Gemini, and AntiGravity) for language polishing, grammar correction, and code refactoring. These tools were used solely to improve the clarity of the authors' original ideas and to support technical implementation. All research claims, interpretations, and final results were generated by the human authors, who reviewed and verified all AI-assisted outputs for accuracy and adherence to ACL ethics policies. We take full responsibility for the final content of the manuscript.