What Makes Retrieval Work for Pedagogical Dialogue Annotation? Indexing Granularity over Retriever Adaptation

Jinsook Lee, Kirk Vanacore, Zhuqian Zhou, Bakhtawar Ahtisham & Rene F. Kizilcec
Cornell University
Ithaca, NY, 14850, USA
{jl3369, kpv27, zz968, ba453, kizilcec}@cornell.edu

Abstract

Annotating tutoring dialogue for pedagogical moves requires trained coders and extensive calibration, which is a major bottleneck that limits the scale of research on tutoring quality. Retrieval-augmented prompting can ease this bottleneck while leaving the generative model frozen, but it is unclear whether the gain comes from the retriever, the segmentation of the dialogue, or the granularity of the exemplar index. We study retrieval-augmented prompting on two English mathematics tutoring datasets (TalkMoves and Eedi) that share a six-category tutor move taxonomy, using three LLM backbones (GPT-5.2, Claude Sonnet 4.6, and Qwen3-32B). On a matched subset where every condition returns a valid label, we find that retrieved demonstrations raise Cohen's κ from 0.41 to 0.55 on TalkMoves and to 0.64 on Eedi for the strongest backbone. Index granularity accounts for most of the improvement and retriever training and segmentation contribute little. Fine-tuning the embedder shifts κ by 0.04 to $+0.05$, lexical BM25 matches dense retrieval to within 0.02 everywhere, and semantic chunking does not beat a three-utterance fixed window. Moving the index unit from the chunk down to the individual utterance, with the parent chunk returned as context, contributes $+0.07$ to $+0.14$. These findings suggest that the unit of the index matters more than the training of the retriever for retrieval-augmented annotation of tutoring dialogue.

1 Introduction

Large language models (LLMs) are increasingly deployed across educational technology for diverse applications, from generating instructional materials and supporting assessment design to real-time tutoring (Scarlatos et al., 2025). While prior work has examined how to adapt or optimize models for specific educational tasks, far less is known about how well LLMs perform when interpreting authentic pedagogical interactions (Chiang & Lee, 2023; He et al., 2024; Anagnostopoulou et al., 2025). This gap is particularly important because educational applications often require understanding nuanced, context-dependent concepts that differ fundamentally from general-purpose language understanding tasks. For example, tutoring moves capture the communicative intent behind tutor and student utterances in educational interactions. A tutor might scaffold a student's reasoning through guided questions, directly provide an answer, press for deeper justification, or manage the conversational flow. Studies demonstrate that certain tutoring moves, such as scaffolding, correlate with deeper conceptual understanding, while excessive telling does not (Michaels et al., 2008; Michaels & O'Connor, 2015; Forbes-Riley & Litman, 2005). Therefore, identifying these moves reliably is essential for evaluating tutoring quality at scale, providing formative feedback to human instructors, and training the reward models that support intelligent tutoring systems. However, tutoring move annotation remains a manual, expert-intensive process. Codebooks such as the TalkMoves framework (Suresh et al., 2022) require trained annotators who understand pedagogical theory, and achieving acceptable inter-rater reliability demands extensive calibration sessions (Vail & Boyer, 2014). This bottleneck limits the scale at which tutoring interactions can be analyzed, a constraint that grows increasingly

untenable as online tutoring platforms generate millions of dialogue sessions per year. The challenge is that tutoring moves are abstract and inherently contextual. The distinction between those dialogue acts depends on surrounding conversational context, the student's prior responses, and the tutor's pedagogical intent rather than the surface form of a single utterance. While LLMs can classify general dialogue acts with reasonable accuracy (Shan et al., 2023; Tan et al., 2023), general-purpose schemas do not capture the pedagogical subtlety required for educational dialogue. We address these limitations with a three-stage approach that decouples domain adaptation from generation. In the first stage, we fine-tune a lightweight sentence embedding model on tutoring dialogue corpora with Multiple Negatives Ranking Loss (MNRL), adapting the representation space so that utterances serving similar pedagogical functions cluster together regardless of surface lexical variation, and use these embeddings to segment dialogues into coherent chunks for FAISS indexing. In the second stage, we retrieve the most semantically similar labeled examples from the index. In the third stage, retrieved examples are presented as few-shot demonstrations alongside the annotation codebook to a general-purpose LLM, which performs classification through in-context learning (ICL) (Dong et al., 2024; Chen et al., 2024). All domain adaptation resides in the retriever that makes the pipeline portable across LLM backbones and reusable as new models become available.

Our contributions are as follows.

1. A **three-stage retrieval-augmented annotation pipeline** that combines domain-adapted embedding, utterance-level indexing with parent chunk context, and codebook-grounded in-context learning with a frozen LLM. Our experiments show relative improvements in Cohen's κ over zero-shot baselines ranging from $+31\%$ to nearly $4\times$ across three LLM backbones on a matched evaluation subset.
2. An **utterance-level indexing strategy** that preserves label-specific signal during search while returning parent chunks as few-shot demonstrations, paired with a dynamic semantic chunking algorithm calibrated against sparse ground truth boundaries.
3. A comprehensive **ablation study** across retrieval conditions, retrieval depth, and LLM backbones on two tutoring dialogue datasets, demonstrating that index granularity, not chunking strategy or embedding quality, accounts for the majority of the performance gain over codebook-only prompting.

2 Related Work

2.1 Dialogue Act Annotation in Educational Settings

Dialogue act annotation has a long history in educational research. The Accountable Talk framework (Michaels et al., 2008) established the theoretical foundation for analyzing classroom discourse through the lens of talk moves that promote productive discussion. Suresh et al. (2022) operationalized this framework computationally by releasing the TalkMoves dataset, which contains K-12 mathematics lesson transcripts annotated at the utterance level for teacher and student discursive moves. Early computational approaches relied on supervised models trained on these labeled corpora (Jensen et al., 2021), with subsequent work addressing class imbalance through AUC optimization (Lin et al., 2023) and integrating multiple annotation perspectives to capture the multi-functionality of tutoring utterances (Naim et al., 2025). A consistent finding is that inter-rater reliability remains a bottleneck which requires extensive calibration even with refined codebooks (Vail & Boyer, 2014).

2.2 Annotating Dialogue Acts with LLMs

Researchers increasingly rely on LLMs to scale annotation, though performance relative to human experts varies by task complexity and workflow design (Su & Ye, 2025; Yu et al., 2024). While LLMs perform well in constrained settings, complex discourse annotation typically requires structured pipelines to accommodate real-world taxonomies and multi-label cases (Ostyakova et al., 2023). A growing body of work frames LLM annotation as a

process design problem, improving robustness through ensembling and orchestration (Na & Feng, 2025; Ahtisham et al., 2026). Particularly in education, some approaches range from using GPT-family models as crowdsourced annotators (He et al., 2024) to generating synthetic tutoring dialogue data to improve coverage of rare codes (Shan et al., 2023) and applying active learning to reduce annotation costs (Tan et al., 2023). Frameworks like EduDCM further improve reliability by decomposing constructs into discrete components such as acts and events, enforcing consistency checks across multiple LLMs (Qi et al., 2025). Despite these advances, LLMs still struggle with multi-party and goal-oriented dialogue act classification where overlapping interactional threads and complex schemas exceed current capabilities (Qamar et al., 2025; Labruna et al., 2023). Comparisons of human and LLM assessment find that models struggle with pedagogically nuanced distinctions without domain-specific guidance (Anagnostopoulou et al., 2025), motivating human-in-the-loop pipelines that integrate verification steps (Tavakoli & Zamani, 2025; Shah et al., 2025).

2.3 Retrieval-Based In-Context Learning and Domain Adaptation

Retrieval-augmented generation (RAG), introduced by Lewis et al. (2020), grounds model outputs in retrieved external evidence at inference time instead of relying on parametric knowledge. This is especially consequential for annotation tasks involving rare or fine-grained categories, where Li et al. (2024) showed selective retrieval yields the largest gains. Chen et al. (2024) extended this to classification via retrieval databases of labeled demonstrations, while Agarwal et al. (2024) and Dong et al. (2024) both highlight that demonstration selection is among the most consequential design choices in ICL. How documents are segmented into retrieval units further shapes performance (Wang et al., 2025), and in tutoring dialogue this is particularly high-stakes since fixed-size windows routinely fragment coherent instructional exchanges.

Segmenting discourse into coherent units also has a long history of its own in the thematic segmentation literature. TextTiling (Hearst, 1997) detects topic boundaries from dips in lexical cohesion between adjacent windows, C99 (Choi, 2000) clusters a sentence-similarity matrix, and later supervised neural segmenters (Koshorek et al., 2018) and topic-aware models such as SECTOR (Arnold et al., 2019) learn boundaries directly from labeled corpora. Our boundary detector (Section 3.2) is best understood as a dense-embedding variant of TextTiling, with cosine similarity between window-averaged sentence embeddings replacing lexical cohesion.

The effectiveness of RAG also depends on the quality of the retrieval embedding space. General-purpose embeddings often fail to capture domain-specific semantic similarity, returning examples that are lexically similar but functionally irrelevant. Adapting the retriever to the target domain through unsupervised LLM adaptation (Liu et al., 2024; BehnamGhader et al., 2024) or self-training (Xu et al., 2025) is often more impactful than adapting the generator.

3 Method

3.1 Task Formulation

Figure 1 shows the overview of the proposed annotation pipeline. Each utterance in a tutoring dialogue has associated text t_i, a speaker role (tutor or student), and optionally a sparse ground-truth label from a predefined codebook. The label set follows the TalkMoves taxonomy with six categories (Suresh et al., 2022), namely Keeping Everyone Together (KET), Getting Students to Relate (GSR), Restating (RES), Pressing for Accuracy (ACC), Revoicing (REV), and Pressing for Reasoning (REA). The task is to predict the label for each tutor utterance by constructing a retrieval index from the corpus, dynamically building context windows around target utterances, and passing both local context and retrieved examples to a frozen LLM for in-context classification.

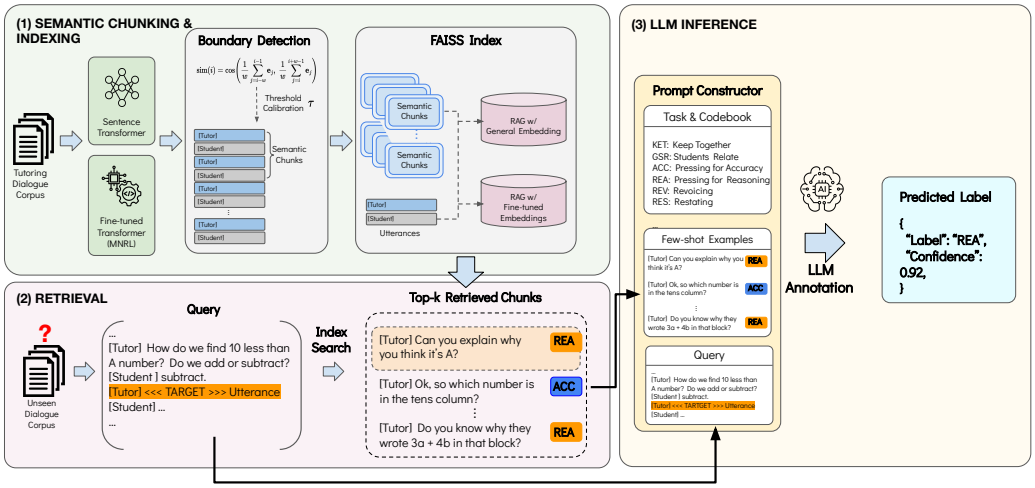

Figure 1: Overview of the proposed RAG-based annotation pipeline.

3.2 Dynamic Semantic Chunking

When applying RAG, setting proper chunking boundaries is important (Wang et al., 2025; Bhat et al., 2025). If a chunk is too small, not enough information would be retrieved and if a chunk is too big, the information would be diluted. Our approach begins by partitioning the dialogue corpus into semantically coherent chunks that preserve label homogeneity while respecting session boundaries. We embed all utterances using a sentence transformer model ϕ such that $\mathbf{e}_i = \phi(t_i)$ where $\mathbf{e}_i \in \mathbb{R}^d$, L2-normalized to enable cosine similarity as a dot product. Semantic boundaries are detected by computing the smoothed cosine similarity between overlapping context windows, a dense-embedding analogue of TextTiling's lexical cohesion score (Hearst, 1997). For each position i in the dialogue, we compute

$$\mathrm{sim}(i) = \cos\left(\frac{1}{w}\sum_{j=i-w}^{i-1}\mathbf{e}_j,\ \frac{1}{w}\sum_{j=i}^{i+w-1}\mathbf{e}_j\right) \tag{1}$$

where w is the window size (default $w = 2$). A low similarity score at position i indicates a candidate semantic boundary where the topic or pedagogical function shifts. To determine the boundary threshold τ, we leverage sparse ground truth annotations. We identify true boundaries as positions where consecutive labeled utterances have differing labels, and non-boundaries as positions where consecutive labeled utterances share the same label. We then sweep τ over $[0.3, 1.0)$ in increments of 0.01 and select the threshold that maximizes F1 on this boundary classification problem. We additionally impose a safety cap where τ cannot exceed the median of all observed similarity scores, preventing over-segmentation when fine-tuned embeddings exhibit representation collapse. Chunks are created by splitting at positions where $\mathrm{sim}(i) < \tau$, respecting a minimum size of 2 and maximum of 20 utterances. The minimum of 2 follows from the window size $w = 2$ in Eq. 1, because a single-utterance chunk has no defined boundary similarity. The maximum of 20 prevents runaway chunks in long stretches where one speaker talks without interruption and the similarity never falls below τ. This upper limit almost never applies. Across the 16,178 training chunks, the median chunk is 5 utterances long, the 95th percentile is 7, the 99th percentile is 10, and only 0.01% of chunks hit the cap.

3.3 Domain-Adapted Embeddings

To improve semantic separation of tutoring dialogue labels in the embedding space, we fine-tune `BGE-large-en-v1.5` on labeled utterances from the TalkMoves and Eedi training sets combined. We combine both datasets because they share the same label taxonomy, and pooling increases the diversity of in-batch negatives for MNRL while improving coverage

of rare labels that are underrepresented in either dataset alone. Training on both classroom and dyadic chat data also encourages the embedding space to capture pedagogical function across interaction formats rather than surface patterns specific to one setting. To prevent data leakage through the embedding geometry, we use only training-split utterances for fine-tuning. Training data consists of (anchor, positive) pairs sampled within each label group, capped at 3,000 pairs per label to prevent frequent labels from dominating the gradient. The cap of 3,000 is a heuristic default rather than a tuned value. Appendix J explains the rationale and reports a sensitivity sweep showing that the pipeline is insensitive to the exact cap. We optimize with MNRL (Henderson et al., 2017), where all other in-batch examples serve as implicit negatives. For a batch of B pairs, the loss for pair i is

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(\mathbf{a}_i, \mathbf{p}_i)/T)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{a}_i, \mathbf{p}_j)/T)} \tag{2}$$

where sim is cosine similarity and T is a temperature parameter. Training proceeds for 3 epochs with batch size 64 and we saved the best checkpoint by evaluation Spearman correlation.

3.4 RAG Index Construction

We construct three FAISS indexes (Johnson et al., 2021): (i) a chunk-level index using Cohere embeddings, (ii) a chunk-level index using our fine-tuned BGE embeddings, and (iii) an utterance-level index using the fine-tuned BGE embeddings, in which each labeled utterance is embedded individually and its parent chunk is retrieved at query time. For the utterance-level index, we retrieve the top k most similar utterances and return their parent chunks as few-shot demonstrations.

MNRL optimizes same-label similarity but does not explicitly enforce separation between different labels, which can cause representation collapse where all embeddings cluster into a narrow cone with pathologically high pairwise similarity. To address this, we apply mean-centering by subtracting the mean embedding vector from all stored embeddings before indexing. The same mean vector is subtracted from query embeddings at search time.

3.5 Query Construction and Prompting

At inference time, we construct a context window for each target utterance that adapts to the dialogue structure. Given a target utterance at position t, we expand backward and forward from t, stopping at positions where $\text{sim}(i) < \tau$ or at session boundaries, up to a maximum of 10 utterances in each direction. For chunk-level indexes, we embed the full context window as the query vector. For utterance-level indexes, we embed only the target utterance itself. Given the query, we search the FAISS index for the top k entries (default $k = 3$). For utterance-level indexes, each retrieved utterance's parent chunk is returned as the demonstration, providing the LLM with full surrounding dialogue context. To avoid data leakage, we exclude the target utterance's own session from retrieval results. The LLM prompt consists of a system message specifying the annotator role, task definition, and full codebook with label definitions, followed by a user message presenting retrieved chunks as labeled examples and the current dialogue context with the target utterance marked by `<<<TARGET>>>`. The model outputs a JSON object with the predicted label and confidence score. Full prompt templates are provided in Appendix A.

4 Experimental Setup

4.1 Datasets

TalkMoves. The TalkMoves dataset[1] contains human-transcribed K-12 mathematics classroom transcripts drawn from whole-class discussions, small-group problem-solving, and

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

online lesson contexts (Suresh et al., 2022). Utterances are annotated with six talk moves organized into three dimensions: *Learning Community* (KET, GSR, RES), *Content Knowledge* (ACC), and *Rigorous Thinking* (REV, REA). We focus exclusively on teacher talk moves. After removing 2 sessions with fewer than 10 human labels, we retain 61 sessions split into 49 training and 12 test (Table 1). The label distribution is heavily imbalanced, with ACC (41%) and KET (27%) dominating while REA and GSR each comprise fewer than 6%.

Eedi. The Eedi dataset (Zent et al., 2025)[2] is drawn from a UK-based online tutoring platform deployed in over 19,000 schools worldwide. Conversations are initiated when a student requests help from an on-demand expert tutor while working on a math lesson, producing dyadic chat-based dialogues anchored to a specific question. The data were collected from November 2021 to February 2025 and annotated with the same six TalkMoves categories used for annotation.

Dataset	Sessions	Utterances	Avg. Utt	Annotated (%)	Train	Test
TalkMoves	61	31,171	511	8,407 (27%)	49	12
Eedi	1,965	46,147	23	30,001 (65%)	1,570	395

Table 1: Dataset statistics. Annotated counts and percentages reflect labels across the full dataset (train and test combined). Avg. Utt is the mean utterances per session.

4.2 Conditions and Implementation

We compare five retrieval conditions (Table 2). The NO_RAG condition provides only the codebook and local dialogue context. The RAG_BM25 condition retrieves chunk-level demonstrations via Okapi BM25 (Robertson & Zaragoza, 2009) over the same corpus with identical session exclusion, providing a lexical-retrieval reference point. The RAG_NO_FINETUNE condition uses Cohere `embed-english-v3.0` for semantic chunking and chunk-level retrieval. The RAG_FINETUNED_CHUNK condition replaces the embedding model with our fine-tuned BGE while retaining chunk-level indexing, isolating the effect of embedding quality from index granularity. The RAG_FINETUNED_UTT condition uses the same fine-tuned BGE model but indexes at the utterance level with parent chunk retrieval. We additionally evaluate fixed-window chunking baselines (Appendix D). To test

Condition	Chunking	Retrieval	Index	Returned
NO_RAG	None	None (codebook only)	–	–
RAG_BM25	Semantic (Cohere)	Okapi BM25 (lexical)	Chunk	Chunk
RAG_NO_FINETUNE	Semantic (Cohere)	Cohere embed-english-v3.0	Chunk	Chunk
RAG_FINETUNED_CHUNK	Semantic (FT-BGE)	Fine-tuned BGE (MNRL)	Chunk	Chunk
RAG_FINETUNED_UTT	Semantic (FT-BGE)	Fine-tuned BGE (MNRL)	Utt.	Chunk

Table 2: Experimental conditions. Index denotes the granularity of FAISS entries (BM25 uses an inverted index over the same chunks). Returned denotes what is passed to the LLM as a few-shot demonstration. RAG_FINETUNED_CHUNK and RAG_FINETUNED_UTT share the same embedding model, isolating the effect of index granularity. Additional ablation conditions are reported in Appendix D.

the generality of the pipeline across LLM backbones, we evaluate the main conditions with three models: GPT-5.2 and Claude Sonnet 4.6, both commercial models accessed via their providers' APIs, and Qwen3-32b, an open-weight model; all three are reached through LiteLLM's unified interface with temperature 1.0, which we retain as the default to avoid artificially constraining the model's output distribution. Lower temperatures may improve consistency but risk suppressing valid alternative predictions for ambiguous utterances. The default retrieval depth is $k = 3$, with ablation over $k \in \{0, 1, 3, 5\}$ for all three models.

[2] https://huggingface.co/datasets/Eedi/Question-Anchored-Tutoring-Dialogues-2k

Our primary metric is Cohen's κ. Because the conditions differ in how often they produce a valid six-class label, and evaluating each condition only on its own valid predictions can inflate metrics for failure-prone conditions, we report main results on a *matched subset*: for each (model, dataset) pair, all conditions are evaluated on the intersection of utterances for which every condition produced a valid label (Section 5.1); full-set results are reported in Appendix I for completeness. We additionally report per-label precision, recall, and F1. Statistical significance is assessed via permutation tests ($n = 2{,}000$) for κ and McNemar's test for accuracy. For embedding fine-tuning, we use `BGE-large-en-v1.5` (Xiao et al., 2023) with batch size 64 for 3 epochs.

5 Results

5.1 Domain-Adapted Retrieval Consistently Improves Annotation Quality Across All Models and Datasets

		TalkMoves		Eedi	
Model	Condition	$N(\cap)$	κ	$N(\cap)$	κ
GPT-5.2	NO_RAG	1,131	0.404	2,623	0.483
GPT-5.2	RAG_BM25	1,131	0.447	2,623	0.547
GPT-5.2	RAG_NO_FINETUNE	1,131	0.439	2,623	0.556
GPT-5.2	RAG_FINETUNED_UTT	1,131	0.530	2,623	0.638
Sonnet 4.6	NO_RAG	977	0.407	2,095	0.407
Sonnet 4.6	RAG_BM25	977	0.443	2,095	0.523
Sonnet 4.6	RAG_NO_FINETUNE	977	0.460	2,095	0.512
Sonnet 4.6	RAG_FINETUNED_UTT	977	**0.547**	2,095	**0.640**
Qwen3-32b	NO_RAG	296	0.282	670	0.147
Qwen3-32b	RAG_BM25	296	0.393	670	0.474
Qwen3-32b	RAG_NO_FINETUNE	296	0.372	670	0.465
Qwen3-32b	RAG_FINETUNED_UTT	296	0.483	670	0.585

Table 3: Matched-subset Cohen's κ across three LLM backbones on both test sets ($k = 3$). All conditions, including RAG_FINETUNED_CHUNK (Appendix H), are scored on the same intersection $N(\cap)$ of utterances where every condition produced a valid label, which removes the differential-coverage confound. Bold marks the best result per dataset. Full-set results appear in Appendix I. In paired permutation tests ($n = 2{,}000$), RAG_FINETUNED_UTT exceeds every other condition in all 18 contrasts ($p < .05$ to $p < .001$), while RAG_BM25 and RAG_NO_FINETUNE never differ significantly.

Table 3 reports matched-subset Cohen's κ for all three models and retrieval conditions with $k = 3$. The ordering NO_RAG $<$ {RAG_BM25, RAG_NO_FINETUNE} $<$ RAG_FINETUNED_UTT holds in every cell. Lexical BM25 retrieval and general-purpose dense retrieval fall within 0.02 of each other everywhere, while domain-adapted utterance-level retrieval beats both by 0.08 to 0.12. The pattern is independent of the backbone's baseline capability, since Qwen3-32b starts weakest ($\kappa = 0.282$ on TalkMoves and 0.147 on Eedi) yet reaches 0.483 and 0.585 with domain-adapted retrieval. The best results are 0.547 on TalkMoves and 0.640 on Eedi, both from Claude Sonnet 4.6. Gains are larger on Eedi across all models, likely because one-on-one tutoring produces more focused exchanges in which utterance-level retrieval finds closer matches.

Matched-subset and full-set values differ by up to 0.132 in either direction (Appendix I): matching raises GPT-5.2's NO_RAG κ while lowering the retrieval conditions for the weaker backbones, and in all six (backbone, dataset) cells the net effect is to shrink the gap between NO_RAG and RAG_FINETUNED_UTT, so the matched subset understates rather than inflates the retrieval effect. Part of what retrieval gives weaker backbones is coverage. Under NO_RAG, Qwen3-32b commits to a valid label on only about a quarter of utterances, compared with roughly 93% under RAG_FINETUNED_UTT. Since Table 3 conditions on every condition producing a valid label, the gains it reports are accuracy gains net of coverage.

5.2 Retrieval Precision Is the Mechanistic Driver of Annotation Gains

	TalkMoves		Eedi	
Condition	Top-1	Any-k	Top-1	Any-k
RAG_BM25	42.4%	69.5%	57.4%	79.8%
RAG_NO_FINETUNE	39.7%	63.4%	52.9%	75.6%
RAG_FINETUNED_UTT	62.0%	83.4%	73.1%	89.9%
Δ (FT-utt − no-FT)	+22.3pp	+20.0pp	+20.2pp	+14.3pp

Table 4: Retrieval label match rate on the full test set ($k = 3$, leave-one-session-out; $N = 1{,}490$ TalkMoves, $N = 3{,}268$ Eedi). Top-1 measures whether the nearest retrieved example shares the target's gold label. Any-k measures whether at least one of the top-k retrieved examples matches. Because this analysis is computed before any LLM is invoked, it is unaffected by differences in LLM prediction coverage. Lexical BM25 retrieval slightly exceeds general-purpose dense retrieval, and both are well below domain-adapted utterance-level retrieval.

To verify that annotation gains stem from improved retrieval quality, Table 4 reports the label match rate between retrieved demonstrations and the target utterance's gold label. With RAG_FINETUNED_UTT, the top-1 match rate reaches 62.0% on TalkMoves and 73.1% on Eedi, compared to 39.7% and 52.9% for RAG_NO_FINETUNE. Notably, lexical BM25 retrieval (42.4% and 57.4% top-1) slightly *exceeds* general-purpose dense retrieval, indicating that general-purpose embeddings offer essentially no advantage over lexical matching on this task; the entire gap to RAG_FINETUNED_UTT comes from domain-adapted embeddings combined with utterance-level indexing. Because this analysis is computed on the full test set before any LLM is invoked, it is free of the coverage confound discussed in Section 5.1 and provides a direct mechanistic explanation for the κ ordering in Table 3. The effect is most pronounced for rare labels: on TalkMoves, the match rate for ACC jumps from 29% to 79%. Conversely, labels where retrieval quality remains low (e.g., RES on Eedi at 0% for both dense conditions) correspond directly to near-zero annotation κ, confirming that retrieval precision drives annotation quality. Detailed per-label analysis is in Appendix E.

5.3 Index Granularity Drives Performance Gains

Two confound-isolation analyses show that the dominant design choice is the granularity of the retrieval index rather than the embedding model or the chunking algorithm.

Because RAG_NO_FINETUNE and RAG_FINETUNED_UTT differ in both embedding model and index granularity, we add RAG_FINETUNED_CHUNK, which pairs the fine-tuned BGE embeddings with chunk-level indexing. On the matched subset with GPT-5.2 it reaches $\kappa = 0.450$ on TalkMoves and 0.514 on Eedi, close to RAG_NO_FINETUNE at 0.439 and 0.556. The jump to RAG_FINETUNED_UTT (0.530 and 0.638) therefore comes from utterance-level indexing. Across all backbones the granularity effect ranges from $+0.070$ to $+0.142$ and exceeds the embedding effect, which ranges from -0.042 to $+0.048$, in every cell (Appendix H).

A second isolation holds the retrieval model constant and compares our semantic chunker against naive fixed windows (Appendix D). Windows of $w = 3$ match or slightly exceed semantic chunking under Cohere retrieval, yet both stay well below RAG_FINETUNED_UTT. Any reasonable segmentation primitive suffices, provided the index is built at the utterance level and the parent chunk is returned as context. Performance also improves through $k = 5$, with the gap between conditions widening as k grows (Appendix G).

5.4 Retrieval Improves Rare and Context-Dependent Annotation Performance

Figure 2 presents normalized confusion matrices for Claude Sonnet 4.6. Without retrieval, the model exhibits a strong bias toward REA on Eedi, assigning it to a substantial fraction of utterances from nearly every other category (e.g., 0.32 of GSR and 0.14 of ACC), while on TalkMoves it over-predicts ACC and shows scattered misclassifications across rare labels.

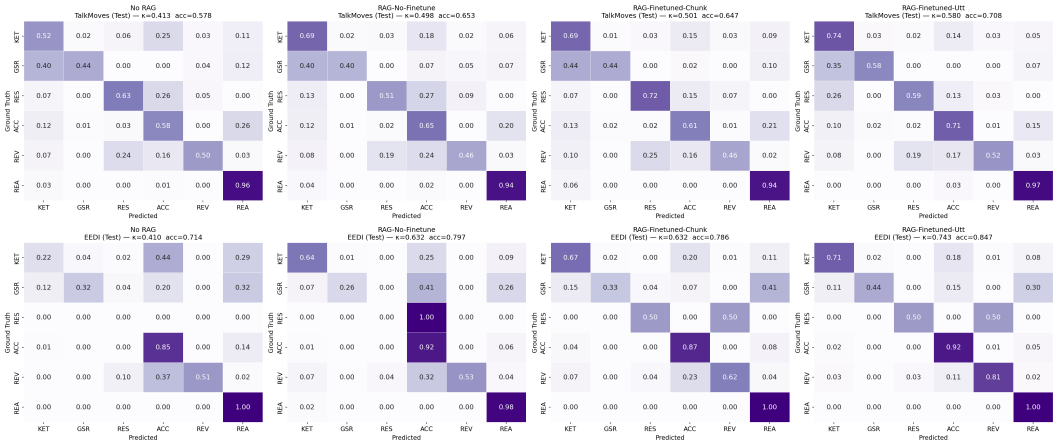

Figure 2: Normalized confusion matrices for TalkMoves test (top) and Eedi test (bottom) with Claude Sonnet 4.6 (full prediction set). Domain-adapted retrieval (RAG_FINETUNED_UTT, right) concentrates mass along the diagonal for all labels. Without retrieval, REA is systematically over-predicted on Eedi.

Retrieval largely eliminates these biases, with RAG_FINETUNED_UTT strengthening diagonal entries across all labels, most notably for REV on Eedi (0.51 to 0.81) and KET on TalkMoves (0.52 to 0.74). The largest absolute gains appear for rare and context-dependent labels: ACC improves from 0.85 to 0.92 and GSR jumps from 0.32 to 0.44 on Eedi, while GSR rises from 0.44 to 0.58 on TalkMoves. These patterns suggest that retrieval helps by providing disambiguating context rather than merely compensating for label scarcity.

Despite these gains, the most persistent confusion on TalkMoves is ACC predicted when the ground truth is KET (0.14 of KET utterances misclassified even with RAG_FINETUNED_UTT), occurring when organizational utterances include question-like forms that superficially resemble accuracy checks. Rare labels also remain challenging, with REA accounting for fewer than 5% of annotations in both datasets. Per-label κ values and full classification reports are in Appendix B; a qualitative analysis of the persistent failure modes, with annotated examples and their retrieved demonstrations, is in Appendix K.

6 Discussion

Utterance-level indexing as the primary lever for improvement. The dominant performance gain comes from utterance-level indexing, and only marginally from embedding quality. Chunk-level embeddings average over multiple utterances and dialogue acts, which dilutes the label-specific signal that drives accurate retrieval. Utterance-level indexing preserves that signal, so the retriever finds demonstrations sharing the target utterance's pedagogical function even when the surface topic differs. Returning the parent chunk at inference time then supplies the surrounding conversational context that the finer index unit gives up, which is why the two components work together. The segmentation algorithm is largely incidental here. Fixed windows of $w = 3$ match or slightly outperform semantic chunking when the retrieval model is held constant (Appendix D), so any reasonable segmentation primitive suffices provided the index is built at the utterance level. This refines the broader principle that adapting the retriever matters more than adapting the generator (Xu et al., 2025; Liu et al., 2024). What carries the gain is where the index is cut, and the training of the embedder is secondary.

Robustness across LLM backbones. The consistent improvement pattern across GPT-5.2, Claude Sonnet 4.6, and Qwen3-32b confirms that our pipeline's gains are not an artifact of a particular model's strengths. Notably, Qwen3-32b has the weakest no-retrieval baseline but benefits the most from domain-adapted retrieval in relative terms, with matched-subset

κ improving nearly 4× on Eedi (from 0.147 to 0.585). Part of this benefit is a coverage gain in addition to an accuracy gain: retrieval induces weaker models to commit to a valid six-class label far more often (from roughly 22–24% of utterances under NO_RAG to roughly 93% under RAG_FINETUNED_UTT for Qwen3-32b), and the matched-subset evaluation in Table 3 shows that large accuracy gains remain even after conditioning coverage out. This suggests that the pipeline is particularly valuable for models with weaker zero-shot classification capabilities, effectively compensating for limited parametric knowledge through high-quality retrieved demonstrations. Since all domain adaptation resides in the 335M-parameter embedding model and FAISS index, the pipeline can be paired with any LLM backend without retraining.

Limitations. Our evaluation is restricted to English-language mathematics tutoring annotated with a single shared taxonomy, and our generality claims are conditional on that taxonomy: the retrieval architecture is taxonomy-dependent, since the fine-tuned embedder and the index store utterances labeled with a specific six-class codebook, so porting to another schema requires re-fine-tuning the embedder. The two datasets do span meaningfully different interaction formats, with TalkMoves capturing multi-party classroom discussions and Eedi consisting of dyadic online chat, and some categories (e.g., Getting Students to Relate) may function differently across these settings. Future work should evaluate generalization across different annotation taxonomies (e.g., STEM tutoring acts, Socratic dialogue codes), subject domains beyond mathematics, and languages beyond English to determine whether domain-adapted retrieval transfers or requires taxonomy-specific re-training. The method also presupposes an existing, expert-annotated seed corpus: domain adaptation of the retriever and the retrieval index itself both require sparse gold labels, so the pipeline scales expert annotation rather than replacing it. In deployment we envision a human-in-the-loop regime in which the annotated seed set is grown under expert review (Appendix F), rather than assumed to be large from the outset. Finally, because valid prediction rates differ across models and conditions, our main results use a matched-subset evaluation that restricts all conditions to a shared set of utterances (Section 5.1); full-set results, which can inflate metrics for failure-prone conditions, are reported in Appendix I for completeness.

7 Conclusion

We presented a retrieval-augmented pipeline for annotating pedagogical dialogue acts at scale. By combining domain-adapted embeddings, dynamic semantic chunking, and codebook-grounded prompting, our system achieves matched-subset Cohen's κ of up to 0.547 on TalkMoves and 0.640 on Eedi across three LLM backbones, substantially outperforming codebook-only baselines without task-specific LLM fine-tuning. The gain comes from where the index is built rather than how the retriever is trained: index granularity accounts for the majority of it, while the embedding model contributes little, and the improvement pattern holds across models of varying capability. Utterance-level indexing with parent chunk retrieval consistently outperforms chunk-level indexing by preserving label-specific signal while providing rich conversational context. Future directions include extending to non-mathematical tutoring domains and other annotation taxonomies, integrating active learning for iterative index improvement, multi-turn annotation that captures dialogue-level pedagogical strategies, and alternative training objectives that enforce inter-label separation without post-hoc corrections.

Acknowledgments

This work was conducted as part of the National Tutoring Observatory, supported by Gates Foundation and Learning Commons.

Ethics Statement

We analyze de-identified tutoring dialogue datasets in accordance with our Institutional Review Board (IRB)-approved protocol (IRB0149250). The TalkMoves dataset was de-

identified by humans before it was open sourced, and the Eedi dataset was de-identified by the tutoring provider with participant consent for research use. We follow IRB-approved procedures for data storage and access and use these data solely to understand and improve dialogue annotation workflows. Re-identification attempts, user profiling, or any use that could enable harm to individual participants is prohibited. Although the data are de-identified, conversational text can carry residual privacy risk. We therefore minimize verbatim excerpts, avoid reporting sensitive attributes, and present results in aggregate wherever possible. Automated annotation of tutoring moves is intended to *augment* rather than replace expert human judgment. The confidence scores produced by our pipeline support a human-in-the-loop workflow in which uncertain predictions are flagged for manual review. We caution against deploying automated labels as ground truth without such verification, particularly for high-stakes decisions such as teacher evaluation or student assessment. Our system relies on commercial LLMs (GPT-5.2, Claude Sonnet 4.6) and an open-weight LLM (Qwen3-32b) accessed via API, which introduces dependence on external services and associated cost, latency, and data-handling considerations; sending tutoring dialogue to third-party APIs also raises data-governance obligations. Researchers applying this pipeline should ensure that any data sent to external APIs complies with applicable data governance policies and the terms under which the dialogue data were collected. Finally, the label taxonomy was developed for mathematics tutoring in English and may not transfer to other languages, subject areas, or cultural contexts without careful adaptation and validation. We used Claude for editorial assistance in drafting and revising portions of this manuscript.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. In *Advances in Neural Information Processing Systems*, 2024.

Bakhtawar Ahtisham, Kirk Vanacore, Jinsook Lee, Zhuqian Zhou, Doug Pietrzak, and Rene F Kizilcec. Ai annotation orchestration: Evaluating llm verifiers to improve the quality of llm annotations in learning analytics. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pp. 447–456, 2026.

Aliki Anagnostopoulou, Nils Feldhus, Yi-Sheng Hsu, Milad Alshomary, Henning Wachsmuth, and Daniel Sonntag. Human and llm-based assessment of teaching acts in expert-led explanatory dialogues. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse*, pp. 166–181, 2025.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A Gers, and Alexander Löser. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184, 2019.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. In *Proceedings of the 1st Conference on Language Modeling*, 2024.

Sinchana Ramakanth Bhat, Max Rudat, Jannis Spiekermann, and Nicolas Flores-Herr. Rethinking chunk size for long-document retrieval: A multi-dataset analysis. *arXiv preprint arXiv:2505.21700*, 2025.

Huiyao Chen, Yu Zhao, Zulong Chen, Mengjia Wang, Liangyue Li, Meishan Zhang, and Min Zhang. Retrieval-style in-context learning for few-shot hierarchical text classification, 2024. URL https://arxiv.org/abs/2406.17534.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870/.

Freddy YY Choi. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, 2000.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Katherine Forbes-Riley and Diane J Litman. Correlating student acoustic-prosodic profiles with student learning in spoken tutoring dialogues. In *Proc. Interspeech 2005*, pp. 157–160, 2005.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 165–190, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-industry.15/.

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

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. 2017. URL https://arxiv.org/abs/1705.00652.

Emily Jensen, Samuel L. Pugh, and Sidney K. D'Mello. A deep transfer learning approach to modeling teacher discourse in the classroom. In *Proceedings of the 11th International Learning Analytics and Knowledge Conference*, pp. 302–312, 2021.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 469–473, 2018.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Hui Xue', and Jun Huang. On the role of long-tail knowledge in retrieval augmented large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 120–126, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.12. URL https://aclanthology.org/2024.acl-short.12/.

Jionghao Lin, Wei Tan, Ngoc Dang Nguyen, David Lang, Lan Du, Wray Buntine, Richard Beare, Guanliang Chen, and Dragan Gašević. Robust educational dialogue act classifiers with low-resource and imbalanced datasets. In *International Conference on Artificial Intelligence in Education*, pp. 114–125. Springer, 2023.

Zheng Liu, Chaofan Li, Shitao Xiao, Yingxia Shao, and Defu Lian. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3490–3500, 2024.

S. Michaels and C. O'Connor. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussions. In Lauren B. Resnick, Christa Asterhan, and Sherice N. Clarke (eds.), *Socializing Intelligence Through Academic Talk and Dialogue*, pp. 333–347. American Educational Research Association, Washington DC, 2015.

Sarah Michaels, Catherine O'Connor, and Lauren B. Resnick. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, 27(4):283–297, 2008. doi: 10.1007/s11217-007-9071-1.

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

Jannatun Naim, Jie Cao, Fareen Tasneem, Jennifer Jacobs, Brent Milne, James Martin, and Tamara Sumner. Towards actionable pedagogical feedback: A multi-perspective analysis of mathematics teaching and tutoring dialogue. In *Proceedings of the 18th International Conference on Educational Data Mining*, 2025.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 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*, pp. 242–254, 2023.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 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)*, pp. 26219–26237, Vienna, Austria, jul 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1271. URL https://aclanthology.org/2025.acl-long.1271/.

Changyong Qi, Longwei Zheng, Yuang Wei, Haoxin Xu, Peiji Chen, and Xiaoqing Gu. Edudcm: a novel framework for automatic educational dialogue classification dataset construction via distant supervision and large language models. *Applied Sciences*, 15(1): 154, 2025.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Alexander Scarlatos, Ryan S. Baker, and Andrew Lan. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 1–10, 2025.

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

Dapeng Shan, Deliang Wang, Chenwei Zhang, Ben Kao, and Carol KK Chan. Annotating educational dialog act with data augmentation in online one-on-one tutoring. In *International Conference on Artificial Intelligence in Education*, pp. 472–477. Springer, 2023.

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

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4654–4662, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.497/.

Wei Tan, Jionghao Lin, David Lang, Guanliang Chen, Dragan Gašević, Lan Du, and Wray Buntine. Does informativeness matter? active learning for educational dialogue act classification. In *International Conference on Artificial Intelligence in Education*, pp. 176–188. Springer, 2023.

Leila Tavakoli and Hamed Zamani. 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, pp. 92–102. Association for Computing Machinery, 2025. doi: 10.1145/3731120.3744574.

Alexandria Katarina Vail and Kristy Elizabeth Boyer. Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes. In *International conference on intelligent tutoring systems*, pp. 199–209. Springer, 2014.

Zhitong Wang, Cheng Gao, Chaojun Xiao, et al. Document segmentation matters for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C. Ho, Carl Yang, and Qi He. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 2025.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 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, 2024.

Matthew Zent, Digory Smith, and Simon Woodhead. PIIvot: A lightweight NLP anonymization framework for question-anchored tutoring dialogues. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27479–27488, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1397. URL https://aclanthology.org/2025.emnlp-main.1397/.

A Prompt Templates

This section presents the full annotation prompt used in our pipeline. The prompt consists of a system message defining the annotator role, codebook, and output format, followed by a user message presenting retrieved examples and the target dialogue context.

A.1 System Message

You are an expert educational discourse analyst. Your task is to label each teacher utterance with one move from the allowed moves list.

Workflow

1. Read the dialogue carefully.
2. For each teacher utterance, assign exactly ONE Move from the Allowed Moves list.
3. If an utterance could fit multiple moves, choose the one that best represents the communicative function in context.
4. If no Move applies (e.g., the utterance is off-topic, evaluative, unclear, or unrelated to the lesson), do not label it and leave it as None.

Allowed Moves & Definitions

- **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.

Rules

- Annotate ONLY the utterance marked with `<<<TARGET>>>`.
- Consider the surrounding context to understand the utterance's function.
- If example annotations from similar dialogues are provided, use them as reference.
- Respond with ONLY a JSON object: `{"label": "<label or null>", "confidence": <0.0-1.0>}`
- If no Move applies, use `null` for label.
- Do not include any other text outside the JSON.

A.2 User Message

For retrieval-augmented conditions, the user message begins with k retrieved examples, each showing the chunk text and its majority label. For NO_RAG, this section is omitted.

Similar annotated examples from the corpus
`Example 1 (labeled: <majority_label>)`
`<chunk text> [... up to k examples ...]`
Current dialogue context `<dynamic_context>` Annotate the utterance marked `<<<TARGET>>>` by `[<target_speaker>]`.
Respond with only the JSON object.

The `<dynamic_context>` field contains the context window from Section 3.5, with the target utterance marked by `<<<TARGET>>>`. Each utterance is formatted as `[speaker] utterance text`.

B Additional Results

Unless otherwise noted, the per-label tables and classification reports in this appendix are computed on each condition's full valid-prediction set, which maximizes per-label support for the rarer categories; the headline cross-condition comparison uses the matched subset (Section 5.1, Appendix I).

B.1 Per-Label Cohen's κ (GPT-5.2)

Tables 5 and 6 report per-label Cohen's κ for all three conditions with GPT-5.2, with labels ordered by frequency.

	ACC	KET	REV	RES	REA	GSR
NO_RAG	0.250	0.472	0.205	0.198	0.232	0.367
RAG_NO_FINETUNE	0.473	0.562	0.362	0.386	0.416	0.459
RAG_FINETUNED_UTT	**0.552**	**0.591**	**0.442**	**0.479**	**0.445**	**0.540**
Support (n)	1,921	1,829	367	288	198	145

Table 5: Per-label Cohen's κ on TalkMoves test ($k = 3$, GPT-5.2, full prediction set).

	ACC	KET	REV	REA	GSR	RES
NO_RAG	0.344	0.584	0.334	0.104	0.106	0.006
RAG_NO_FINETUNE	0.551	0.671	0.263	0.250	0.213	0.020
RAG_FINETUNED_UTT	**0.662**	**0.739**	**0.642**	**0.380**	**0.418**	**0.444**
Support (n)	5,286	2,779	1,531	157	81	9

Table 6: Per-label Cohen's κ on Eedi test ($k = 3$, GPT-5.2, full prediction set). RES has only 9 support instances. RAG_NO_FINETUNE underperforms NO_RAG on REV (0.263 vs. 0.334).

B.2 Classification Reports (GPT-5.2)

	Precision	Recall	F1	Support
KET	0.68	0.66	0.67	1,829
GSR	0.55	0.29	0.38	145
RES	0.22	0.30	0.25	287
ACC	0.62	0.41	0.49	1,921
REV	0.56	0.14	0.23	366
REA	0.17	0.96	0.29	198
Accuracy			0.50	4,746
Macro avg	0.47	0.46	0.39	4,746
Weighted avg	0.59	0.50	0.52	4,746
Cohen's $\kappa = 0.315$				

Table 7: NO_RAG classification report on TalkMoves test (GPT-5.2, full prediction set).

C Model Comparison Per-Label Results

Tables 13 and 14 report per-label Cohen's κ for all three LLM backbones under each retrieval condition ($k = 3$, full prediction set).

	Precision	Recall	F1	Support
KET	0.77	0.66	0.71	701
GSR	0.70	0.36	0.47	73
RES	0.46	0.39	0.42	111
ACC	0.68	0.71	0.69	774
REV	0.61	0.29	0.39	147
REA	0.30	0.99	0.46	84
Accuracy			0.64	1,890
Macro avg	0.59	0.57	0.52	1,890
Weighted avg	0.68	0.64	0.64	1,890
Cohen's $\kappa = 0.479$				

Table 8: RAG_NO_FINETUNE classification report on TalkMoves test (GPT-5.2, full prediction set).

	Precision	Recall	F1	Support
KET	0.83	0.65	0.73	534
GSR	0.64	0.49	0.55	43
RES	0.53	0.48	0.50	69
ACC	0.72	0.78	0.75	589
REV	0.68	0.36	0.47	90
REA	0.32	0.97	0.48	62
Accuracy			0.69	1,387
Macro avg	0.62	0.62	0.58	1,387
Weighted avg	0.73	0.69	0.69	1,387
Cohen's $\kappa = 0.539$				

Table 9: RAG_FINETUNED_UTT classification report on TalkMoves test (GPT-5.2, full prediction set).

D Fixed-Window Chunking Ablation

To isolate the contribution of semantic boundary detection from other components of the pipeline, we compare our calibrated semantic chunking against naive fixed-window baselines. Fixed-window chunking partitions each dialogue into non-overlapping segments of exactly w utterances, with no boundary detection or similarity computation. We test $w = 3$ and $w = 5$, both using Cohere embeddings for chunk-level retrieval to match the RAG_NO_FINETUNE condition.

D.1 Overall Results

Table 15 presents Cohen's κ across all chunking conditions on both test sets with $k = 3$ and GPT-5.2. Several observations emerge. First, fixed-window chunking with $w = 3$ matches or slightly outperforms semantic chunking when the same retrieval model (Cohere) is used. On TalkMoves, RAG_FIXED_W3 ($\kappa = 0.500$) exceeds RAG_NO_FINETUNE (0.474). On Eedi, the same pattern holds (0.571 vs. 0.505). This suggests that for chunk-level retrieval with general-purpose embeddings, shorter fixed windows may actually produce better retrieval units than semantic chunking, possibly because smaller chunks reduce the dilution effect discussed in Section 3.4. Second, $w = 3$ consistently outperforms $w = 5$ on both datasets, reinforcing that shorter chunks produce more focused retrieval matches when using chunk-level indexing. Third, and most importantly, RAG_FINETUNED_UTT outperforms all other conditions by a substantial margin despite using semantic rather than fixed-window chunking. This confirms that the dominant performance driver is the combination of domain-adapted embeddings with utterance-level indexing, not the chunking strategy itself. The chunking strategy matters at the margin when all else is held

	Precision	Recall	F1	Support
KET	0.79	0.59	0.67	2,335
GSR	0.21	0.07	0.11	81
RES	0.00	0.17	0.01	6
ACC	0.75	0.65	0.69	5,265
REV	0.95	0.23	0.37	1,334
REA	0.07	0.99	0.13	157
Accuracy			0.57	9,178
Macro avg	0.46	0.45	0.33	9,178
Weighted avg	0.77	0.57	0.63	9,178
Cohen's $\kappa = 0.351$				

Table 10: NO_RAG classification report on Eedi test (GPT-5.2, full prediction set).

	Precision	Recall	F1	Support
KET	0.86	0.67	0.75	1,788
GSR	0.40	0.15	0.22	54
RES	0.01	0.50	0.02	4
ACC	0.78	0.86	0.81	3,522
REV	0.94	0.17	0.29	842
REA	0.16	1.00	0.27	105
Accuracy			0.71	6,315
Macro avg	0.52	0.56	0.39	6,315
Weighted avg	0.81	0.71	0.71	6,315
Cohen's $\kappa = 0.505$				

Table 11: RAG_NO_FINETUNE classification report on Eedi test (GPT-5.2, full prediction set).

equal, but the choice of embedding model and index granularity accounts for the majority of the performance gap.

D.2 Per-Label κ for Fixed-Window Conditions

Tables 16 and 17 break down the fixed-window results by label. An interesting pattern is that RAG_FIXED_W3 achieves notably high κ on RES (0.731) and REV (0.549) on TalkMoves, outperforming even RAG_FINETUNED_UTT on those two labels. This may reflect that Restating and Revoicing are highly localized acts (a tutor repeating or rephrasing what a student just said), making short fixed windows effective retrieval units for these specific categories. However, RAG_FINETUNED_UTT achieves more balanced performance across all labels and substantially higher κ on the labels that require broader contextual understanding (GSR, ACC, REA). On Eedi, RAG_FINETUNED_UTT dominates on every label.

E Retrieval Label Match Rate

To directly evaluate whether domain-adapted embeddings improve retrieval quality independently of LLM annotation, we measure the label match rate between retrieved demonstrations and the target utterance's gold label. For each target utterance, we check whether the top-1 retrieved example shares the same gold label (Top-1) and whether any of the top-k retrieved examples share the label (Any-k, with $k = 3$). This analysis is computed on the full test set ($N = 1{,}490$ TalkMoves, $N = 3{,}268$ Eedi) before any LLM is invoked, so it is unaffected by differences in LLM prediction coverage across conditions.

Domain-adapted retrieval produces a 14 to 22 percentage point improvement in label match rate across both datasets and both metrics. Lexical BM25 retrieval over the same chunks (Table 4) slightly exceeds general-purpose dense retrieval (42.4% vs. 39.7% top-1 on TalkMoves; 57.4% vs. 52.9% on Eedi), indicating that general-purpose embeddings provide essentially

	Precision	Recall	F1	Support
KET	0.96	0.69	0.80	852
GSR	0.44	0.41	0.42	27
RES	0.29	1.00	0.44	2
ACC	0.82	0.93	0.87	1,753
REV	0.95	0.52	0.67	359
REA	0.25	1.00	0.40	52
Accuracy			0.81	3,045
Macro avg	0.62	0.76	0.60	3,045
Weighted avg	0.86	0.81	0.81	3,045
Cohen's $\kappa = 0.659$				

Table 12: RAG_FINETUNED_UTT classification report on Eedi test (GPT-5.2, full prediction set).

Model	Condition	KET	GSR	RES	ACC	REV	REA
GPT-5.2	NO_RAG	0.472	0.367	0.198	0.250	0.205	0.232
GPT-5.2	RAG_NO_FINETUNE	0.562	0.459	0.386	0.473	0.362	0.416
GPT-5.2	RAG_FINETUNED_UTT	0.591	**0.540**	0.479	0.552	0.442	0.445
Sonnet 4.6	NO_RAG	0.421	0.506	0.410	0.397	0.548	0.353
Sonnet 4.6	RAG_NO_FINETUNE	0.577	0.464	0.461	0.480	0.522	0.360
Sonnet 4.6	RAG_FINETUNED_UTT	**0.631**	0.528	**0.542**	**0.594**	**0.566**	**0.448**
Qwen3-32b	NO_RAG	0.393	0.406	0.366	0.090	0.630	0.224
Qwen3-32b	RAG_NO_FINETUNE	0.490	0.327	0.380	0.440	0.359	0.427
Qwen3-32b	RAG_FINETUNED_UTT	0.584	0.411	0.367	0.584	0.405	0.426

Table 13: Per-label Cohen's κ on TalkMoves test across three LLM backbones ($k = 3$, full prediction set). Bold indicates the best RAG_FINETUNED_UTT result per label.

Model	Condition	KET	GSR	RES	ACC	REV	REA
GPT-5.2	NO_RAG	0.584	0.106	0.006	0.344	0.334	0.104
GPT-5.2	RAG_NO_FINETUNE	0.671	0.213	0.020	0.551	0.263	0.250
GPT-5.2	RAG_FINETUNED_UTT	0.739	**0.418**	$0.444^{\dagger}$	0.662	0.642	0.380
Sonnet 4.6	NO_RAG	0.303	0.300	-	0.478	0.643	0.207
Sonnet 4.6	RAG_NO_FINETUNE	0.683	0.292	-	0.670	0.653	0.330
Sonnet 4.6	RAG_FINETUNED_UTT	0.749	0.387	0.104	**0.775**	**0.855**	0.361
Qwen3-32b	NO_RAG	0.201	0.275	-	0.129	0.571	0.088
Qwen3-32b	RAG_NO_FINETUNE	0.690	0.273	-	0.630	0.356	0.443
Qwen3-32b	RAG_FINETUNED_UTT	**0.764**	0.368	0.092	0.734	0.665	**0.400**

Table 14: Per-label Cohen's κ on Eedi test across three LLM backbones ($k = 3$, full prediction set). Bold indicates the best RAG_FINETUNED_UTT result per label. †RES has only 9 support instances. No results were noted with dashes.

no advantage over lexical matching for this task; the full gap to RAG_FINETUNED_UTT is attributable to domain-adapted embeddings combined with utterance-level indexing. The gains are especially pronounced for labels that are rare or context-dependent. On TalkMoves, the any-k match rate for ACC jumps from 29% to 79% with fine-tuned embeddings, directly explaining the large annotation κ improvement for that label. KET, the most frequent label, achieves near-perfect retrieval (96–97% any-k) under both conditions, consistent with its high annotation κ across all models. Conversely, labels where retrieval quality remains low correspond to persistently low annotation performance. RES on Eedi has a 0% match rate for both dense conditions, reflecting its extreme rarity in the training index (9 labeled instances). Neither embedding model can retrieve relevant RES demonstrations, which directly explains the near-zero κ for this label. GSR and REV both remain below 45% any-k

Condition	Chunking	Retrieval	TM κ	Eedi κ
NO_RAG	None	None	0.313	0.351
RAG_NO_FINETUNE	Semantic (Cohere)	Cohere	0.474	0.505
RAG_FIXED_W3	Fixed (w=3)	Cohere	0.500	0.571
RAG_FIXED_W5	Fixed (w=5)	Cohere	0.475	0.519
RAG_FINETUNED_UTT	Fine-tuned BGE	Fine-tuned BGE	**0.539**	**0.659**

Table 15: Fixed-window vs. semantic chunking ablation on test sets ($k = 3$, GPT-5.2, full prediction set). All conditions except NO_RAG and RAG_FINETUNED_UTT use Cohere for retrieval. Sample sizes differ slightly across conditions due to index coverage, which is why the NO_RAG and RAG_NO_FINETUNE values here differ marginally from Appendix I (0.313 vs. 0.315; 0.474 vs. 0.479).

	KET	GSR	RES	ACC	REV	REA
RAG_FIXED_W3	0.554	0.254	**0.731**	0.470	0.549	0.350
RAG_FIXED_W5	0.537	0.396	0.691	0.447	**0.553**	0.293
RAG_FINETUNED_UTT	**0.591**	**0.540**	0.479	**0.552**	0.442	**0.445**

Table 16: Per-label κ on TalkMoves test for fixed-window conditions vs. RAG_FINETUNED_UTT.

	KET	GSR	RES	ACC	REV	REA
RAG_FIXED_W3	0.676	0.195	-0.000	0.599	0.534	0.256
RAG_FIXED_W5	0.607	0.069	0.166	0.552	0.592	0.187
RAG_FINETUNED_UTT	**0.739**	**0.418**	**0.444**	**0.662**	**0.642**	**0.380**

Table 17: Per-label κ on Eedi test for fixed-window conditions vs. RAG_FINETUNED_UTT.

match rate, consistent with their lower support and higher chunk impurity. These patterns confirm that retrieval precision is the mechanistic bottleneck for annotation quality and that the pipeline degrades predictably for labels that are underrepresented in the training index.

F Confidence Calibration Analysis

Our pipeline produces a confidence score alongside each predicted label. We analyze whether these scores are informative by examining (1) how confidence distributions shift across retrieval conditions and (2) whether high-confidence predictions are reliably correct. Figure 3 shows that retrieval shifts the confidence distribution rightward for all three models.

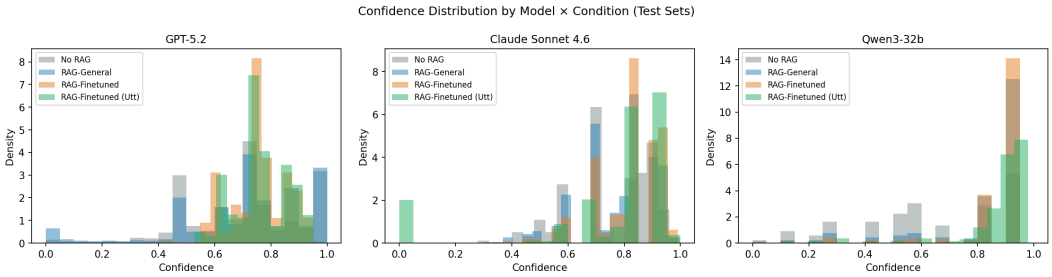

Figure 3: Confidence score distributions by model and retrieval condition on the combined test sets. Domain-adapted retrieval (RAG_FINETUNED_UTT) concentrates prediction mass at higher confidence levels across all three models.

With RAG_FINETUNED_UTT, predictions concentrate in the 0.8–1.0 range, consistent with the

high retrieval label match rates reported in Appendix E. Without retrieval, confidence scores are more dispersed, and for GPT-5.2 a notable mass appears near zero, indicating frequent uncertainty. To assess whether confidence scores support reliable triage, we compute Cohen's κ and coverage on the subset of predictions with confidence $\geq$ 0.9 (Table 18). Two patterns emerge. First, RAG_FINETUNED_UTT substantially improves the reliability of

Model	Condition	κ ($\geq$ 0.9)	Coverage
GPT-5.2	NO_RAG	0.086	19.1%
GPT-5.2	RAG_NO_FINETUNE	0.412	23.8%
GPT-5.2	RAG_FINETUNED_UTT	**0.939**	11.6%
Sonnet 4.6	NO_RAG	0.594	3.9%
Sonnet 4.6	RAG_NO_FINETUNE	0.838	24.3%
Sonnet 4.6	RAG_FINETUNED_UTT	0.896	34.3%
Qwen3-32b	NO_RAG	0.212	17.0%
Qwen3-32b	RAG_NO_FINETUNE	0.611	59.8%
Qwen3-32b	RAG_FINETUNED_UTT	0.685	68.3%

Table 18: Cohen's κ and coverage for predictions with confidence $\geq$ 0.9. Coverage is the fraction of test utterances where the model's confidence exceeds the threshold. GPT-5.2 with RAG_FINETUNED_UTT achieves near-perfect agreement ($\kappa = 0.939$) on its high-confidence subset, though coverage is limited to 11.6%.

high-confidence predictions across all models. GPT-5.2 with RAG_FINETUNED_UTT achieves $\kappa = 0.939$ on its confident subset, compared to 0.086 without retrieval, indicating that domain-adapted retrieval nearly eliminates errors among predictions the model is most certain about. Second, models differ in how they distribute confidence. Sonnet 4.6 with RAG_FINETUNED_UTT achieves both high κ (0.896) and high coverage (34.3%), making it the strongest candidate for a human-in-the-loop workflow where confident predictions are auto-accepted and uncertain ones are flagged for manual review. Qwen3-32b produces the highest coverage (68.3%) but at lower κ (0.685), suggesting it is overconfident relative to its actual accuracy. These results support using confidence scores as a practical triage mechanism in deployment, though the coverage-accuracy tradeoff varies across models and should be calibrated for specific use cases.

G Effect of Retrieval Depth

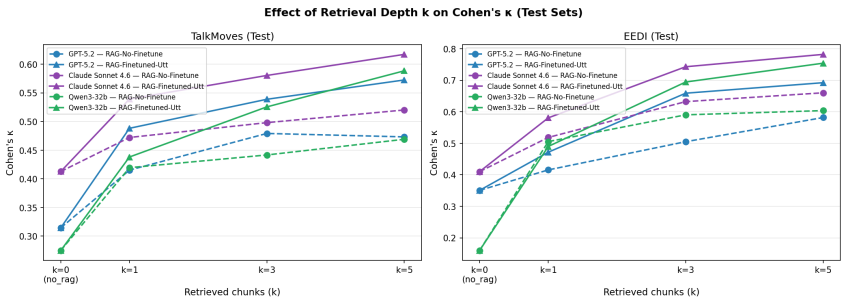

Figure 4: Cohen's κ as a function of retrieval depth k across three LLM backbones on TalkMoves test (top) and Eedi test (bottom). Left panels show RAG_NO_FINETUNE, right panels show RAG_FINETUNED_UTT. Domain-adapted retrieval consistently outperforms general-purpose retrieval at every k for all models.

Figure 4 shows annotation quality as a function of retrieval depth k across all three models. RAG_FINETUNED_UTT outperforms RAG_NO_FINETUNE at every value of k for all models, confirming that domain-adapted retrieval consistently produces more informative demonstrations regardless of the LLM backbone. A single retrieved example ($k = 1$) is clearly

insufficient across all conditions. Performance continues to improve through $k = 5$ on Eedi for all models, with Sonnet 4.6 reaching $\kappa = 0.782$. On TalkMoves, κ also improves through $k = 5$ under RAG_FINETUNED_UTT for all models, while RAG_NO_FINETUNE plateaus or slightly declines after $k = 3$, possibly because additional chunks from a general-purpose index introduce noise. The widening gap between RAG_FINETUNED_UTT and RAG_NO_FINETUNE as k increases suggests that domain-adapted embeddings become increasingly important as retrieval depth grows.

H Confound Isolation Analysis

The comparison between RAG_NO_FINETUNE and RAG_FINETUNED_UTT conflates two changes: embedding model (Cohere vs. fine-tuned BGE) and index granularity (chunk vs. utterance). To disentangle these factors, we evaluate RAG_FINETUNED_CHUNK, which uses the fine-tuned BGE model with chunk-level indexing. Each step changes exactly one variable, and we report results for all three LLM backbones (Tables 19 and 20). κ values in this appendix are computed on the same matched evaluation subset as Table 3.

Model	Condition	κ	Emb $\Delta\kappa$	Gran $\Delta\kappa$
GPT-5.2	RAG_NO_FINETUNE	0.439		
GPT-5.2	RAG_FINETUNED_CHUNK	0.450	+0.011	
GPT-5.2	RAG_FINETUNED_UTT	**0.530**		+0.080
Sonnet 4.6	RAG_NO_FINETUNE	0.460		
Sonnet 4.6	RAG_FINETUNED_CHUNK	0.477	+0.017	
Sonnet 4.6	RAG_FINETUNED_UTT	**0.547**		+0.070
Qwen3-32b	RAG_NO_FINETUNE	0.372		
Qwen3-32b	RAG_FINETUNED_CHUNK	0.341	−0.031	
Qwen3-32b	RAG_FINETUNED_UTT	**0.483**		+0.142

Table 19: Confound isolation on TalkMoves test ($k = 3$, matched evaluation subset). Emb $\Delta\kappa$ is the effect of switching embeddings (Cohere $\rightarrow$ FT-BGE) at chunk-level granularity. Gran $\Delta\kappa$ is the effect of switching granularity (chunk $\rightarrow$ utterance) with FT-BGE.

Model	Condition	κ	Emb $\Delta\kappa$	Gran $\Delta\kappa$
GPT-5.2	RAG_NO_FINETUNE	0.556		
GPT-5.2	RAG_FINETUNED_CHUNK	0.514	−0.042	
GPT-5.2	RAG_FINETUNED_UTT	**0.638**		+0.124
Sonnet 4.6	RAG_NO_FINETUNE	0.512		
Sonnet 4.6	RAG_FINETUNED_CHUNK	0.560†	+0.048	
Sonnet 4.6	RAG_FINETUNED_UTT	**0.640**		+0.080
Qwen3-32b	RAG_NO_FINETUNE	0.465		
Qwen3-32b	RAG_FINETUNED_CHUNK	0.493	+0.028	
Qwen3-32b	RAG_FINETUNED_UTT	**0.585**		+0.092

Table 20: Confound isolation on Eedi test ($k = 3$, matched evaluation subset). †Sonnet RAG_FINETUNED_CHUNK is based on an incomplete run ($\sim$253 records short). Embedding model effect (Emb $\Delta\kappa$) ranges from -0.042 to $+0.048$. Granularity effect (Gran $\Delta\kappa$) ranges from $+0.070$ to $+0.142$.

Across all six model-dataset combinations, switching the embedding model from Cohere to fine-tuned BGE while retaining chunk-level indexing yields small, sign-mixed $\Delta\kappa$ (range: -0.042 to $+0.048$). In contrast, switching from chunk-level to utterance-level indexing with the same fine-tuned BGE model consistently produces large gains (range: $+0.070$ to $+0.142$), exceeding the embedding effect in every cell. This confirms that utterance-level indexing is the primary driver of annotation improvement across all LLM backbones. Chunk-level embeddings average over multiple utterances that may serve different pedagogical functions, diluting the label-specific signal needed for accurate retrieval. Utterance-level

indexing preserves this signal while still returning the full parent chunk as context for the LLM.

I Full-Set Results

For completeness, Table 21 reports Cohen's κ and accuracy computed on each condition's full valid-prediction set, i.e., every utterance for which that condition produced a valid six-class label. Because conditions differ in how often they produce a valid label (e.g., Qwen3-32b commits to a valid label on only 22–24% of utterances under NO_RAG versus roughly 93% under RAG_FINETUNED_UTT), these values are not directly comparable across conditions and can inflate metrics for conditions that fail on harder examples; the matched-subset evaluation in Section 5.1 is the headline comparison. Matched and full-set values differ by at most 0.132 in any cell, but not in a single direction. Matching raises GPT-5.2's NO_RAG κ (+0.089 on TalkMoves, +0.132 on Eedi) while lowering the retrieval conditions for Sonnet 4.6 and Qwen3-32b on Eedi (down to −0.125). Both effects narrow the NO_RAG-to-RAG_FINETUNED_UTT gap, which shrinks under matching in all six (backbone, dataset) cells; the matched subset is therefore the conservative comparison. RAG_BM25 is omitted from this table, and so from the comparison above, as it was run only under the matched protocol.

Model	Condition	TM κ	TM Acc	Eedi κ	Eedi Acc
GPT-5.2	NO_RAG	0.315	0.500	0.351	0.575
GPT-5.2	RAG_NO_FINETUNE	0.479	0.639	0.505	0.709
GPT-5.2	RAG_FINETUNED_CHUNK	0.454	0.620	0.530	0.740
GPT-5.2	RAG_FINETUNED_UTT	0.539	0.688	0.659	0.810
Sonnet 4.6	NO_RAG	0.413	0.578	0.410	0.714
Sonnet 4.6	RAG_NO_FINETUNE	0.498	0.653	0.632	0.797
Sonnet 4.6	RAG_FINETUNED_CHUNK	0.501	0.647	0.632	0.780
Sonnet 4.6	RAG_FINETUNED_UTT	**0.580**	**0.708**	**0.743**	**0.847**
Qwen3-32b	NO_RAG	0.275	0.404	0.160	0.354
Qwen3-32b	RAG_NO_FINETUNE	0.442	0.617	0.590	0.769
Qwen3-32b	RAG_FINETUNED_CHUNK	0.428	0.585	0.545	0.725
Qwen3-32b	RAG_FINETUNED_UTT	0.526	0.667	0.694	0.818

Table 21: Full-set Cohen's κ and accuracy across three LLM backbones on both test sets ($k = 3$), where each condition is evaluated on its own valid-prediction subset. Bold indicates the best result per metric. All RAG vs. NO_RAG differences are significant at $p < .001$ (permutation test, $n = 2{,}000$) except Qwen3-32b κ on TalkMoves for RAG_NO_FINETUNE ($p < .01$).

J Pair-Cap Sensitivity Sweep

The per-label cap of 3,000 (anchor, positive) pairs used for MNRL fine-tuning (Section 3.3) is a heuristic default. To assess sensitivity, we re-fine-tuned the embedding model at alternative cap values under identical conditions (same seed, hardware, and library versions, trained back to back) and measured retrieval label match rate (Table 22). Match rates show no systematic effect of the cap: Eedi rates are essentially constant, and the TalkMoves spread (49.8–55.2% top-1) is non-monotonic in the cap, indicating that fine-tuning stochasticity, not the cap value, drives the variation on the smaller test set. The cap does bind differently per label: 2,000 is the only swept value that crosses below REV's training support of 2,617, truncating a non-dominant label, which motivated 3,000 as the smallest value preserving the full support of all four non-dominant labels. We note that the production embedding model used throughout the paper (cap = 3,000) was trained in an earlier environment and achieves higher absolute match rates on TalkMoves (62.0% top-1); given the controlled sweep shows no cap effect, we attribute that difference to training-environment and run-to-run variance

Cap	TM Top-1	TM Any-k	Eedi Top-1	Eedi Any-k
2,000	55.2%	80.5%	70.0%	87.1%
3,000	49.8%	73.7%	69.0%	87.2%
5,000	53.9%	80.5%	69.9%	89.6%

Table 22: Retrieval label match rate under different per-label pair caps ($k = 3$), with all three models fine-tuned under identical conditions (same seed, hardware, and library versions, trained back to back). On Eedi, rates are flat across caps (within 1.0pp Top-1). On the smaller TalkMoves test set, values vary non-monotonically across caps, consistent with fine-tuning run-to-run variance rather than a cap effect.

rather than to the cap value, and report the sweep from identically-trained models to isolate the variable of interest.

The κ-level sweep (Table 23) confirms cap insensitivity end to end: annotating both full test sets with all three cap models across all three LLM backbones, matched-subset κ varies by at most 0.013 across caps within any (backbone, dataset) cell. Notably, downstream κ is also robust to the retrieval-level variance visible in Table 22: despite their lower TalkMoves match rates, the identically-trained cap models reach κ within 0.005–0.046 of the production model's RAG_FINETUNED_UTT results, indicating that annotation quality saturates well below the production model's retrieval precision.

		TalkMoves κ		Eedi κ	
Model	Cap	Full	Matched	Full	Matched
GPT-5.2	2,000	0.541	0.543	0.667	0.680
GPT-5.2	3,000	0.536	0.543	0.652	0.674
GPT-5.2	5,000	0.536	0.542	0.666	0.677
Sonnet 4.6	2,000	0.574	0.578	0.733	0.735
Sonnet 4.6	3,000	0.574	0.586	0.726	0.729
Sonnet 4.6	5,000	0.570	0.575	0.731	0.734
Qwen3-32b	2,000	0.531	0.534	0.648	0.661
Qwen3-32b	3,000	0.541	0.536	0.654	0.661
Qwen3-32b	5,000	0.526	0.536	0.645	0.648

Table 23: κ-level pair-cap sweep ($k = 3$): each identically-trained cap model used for utterance-level retrieval, annotated by all three backbones. "Matched" restricts to the intersection of valid predictions across the three caps within each (model, dataset). Maximum spread across caps in any cell is 0.013.

K Qualitative Error Analysis

This appendix examines the three persistent failure modes identified in Section 5, each illustrated with brief annotated examples showing minimal dialogue context, the model's prediction, and the gold label. In line with our data-governance commitments (Ethics Statement), excerpts are kept as short as possible and student turns are paraphrased or truncated.

REA over-prediction on Eedi under NO_RAG. Without retrieval, models systematically over-assign Pressing for Reasoning on Eedi (e.g., 0.32 of GSR and 0.14 of ACC utterances are misclassified as REA by Sonnet 4.6). The pattern is that *any* question inviting a substantive response is read as a reasoning press by default; the codebook's narrower sense of REA (pressing students to explain or justify their thinking) is only recovered once retrieved exemplars illustrate the boundary.

- *"So, how can we work out what each section is worth?"* (following the tutor establishing that a number line is divided into 10 sections) — gold **ACC**, predicted REA. The question solicits a mathematical contribution, not a justification of prior reasoning.
- *"There's more than one way you can solve it. What did you do first?"* — gold **ACC**, predicted REA. Eliciting the student's calculation steps is coded as pressing for a mathematical contribution.
- *"Nice! Does that answer make the statement true?"* (after the student supplies an answer) — gold **KET**, predicted REA. A check that the student is following the task logic, not a press for justification.
- *"Does the sequence go up by the same amount each time?"* (immediately after displaying the sequence) — gold **KET**, predicted REA. An orienting yes/no check misread as a reasoning press because of its inferential surface form.

ACC predicted for KET on TalkMoves, surviving RAG_FINETUNED_UTT. Even with domain-adapted retrieval, 0.14 of KET utterances are misclassified as ACC. These are organizational or orienting utterances whose surface content is mathematical, so they superficially resemble accuracy presses:

- *"Is one-tenth equal to 10 one-hundredths?"* (after students confirm they are ready to proceed) — gold **KET**, predicted ACC. In context the teacher is re-orienting the whole class to a shared fact before moving on, not eliciting a new contribution.
- *"Are there any tenths in two-hundredths?"* (guiding a student through place-value columns just discussed) — gold **KET**, predicted ACC.
- *"Did you forget to add 2 zeros?"* (after a student compares their answer with a peer's) — gold **KET**, predicted ACC. A managing move keeping the student aligned with the worked example, phrased as a content question.

These cases sit near the annotation boundary itself: distinguishing them requires the classroom-level context of what the teacher is orchestrating, which neither the local context window nor utterance-level retrieval fully captures.

RES on Eedi: an upstream retrieval failure. RES on Eedi has near-zero annotation κ under most conditions. The retrieval label-match analysis (Appendix E) shows the failure occurs upstream of the LLM: with only 9 labeled RES instances in the training index, no relevant demonstrations can be retrieved, and the model receives no exemplar signal distinguishing verbatim restating from revoicing.

- Student: *"I know the rectangle is 40m"* → Tutor: *"Yes, 40 [thumbs-up]"* — gold **RES** (verbatim repetition of the student's value). Sonnet 4.6 predicts REV; the distinction between repeating word-for-word (RES) and rephrasing with added content (REV) hinges on comparing against the student's prior turn.
- Student: *"Ok is it a term?"* → Tutor: *"No it's not a term"* — gold **RES** (echoing the student's formulation in the negative). All backbones abstain or misclassify; no RES demonstration is available to anchor the decision.
- Student: *"Er 4 over 10?"* → Tutor: *"4/10"* — gold **RES**, correctly predicted by GPT-5.2 and Sonnet 4.6: the bare verbatim echo is the one surface pattern recognizable without demonstrations.

The pipeline thus degrades predictably when a label is absent from the retrieval index; expanding index coverage for rare labels (for example with synthetic exemplars, cf. Shan et al., 2023) is the natural remedy.