

Retrieval-Conditioned Counterfactual Consistency for Classroom Discourse Analysis

-- A controlled study of sparse evidence selection and prediction stability

Emily Garcia *, Hazel Xu

Steinhardt School of Culture, Education, and Human Development, New York University, New York, United States

Corresponding author: Emily Garcia (Email: s9583921@nyu.edu)

Abstract: Classroom discourse analysis depends on detecting pedagogically meaningful reasoning from turns whose relevant context may occur far earlier. This study evaluates a small-footprint retrieval-conditioned classifier augmented with an explicit counterfactual consistency penalty. We manually curated 136 publicly displayed, human-labeled student-response records from the TalkMoves corpus (47 labeled as providing evidence or reasoning). A controlled benchmark inserts each record's original paired context among 8, 32, or 64 distractor turns sampled from the same public excerpt pool; these reconstructed histories are synthetic, whereas the classification labels are authentic. A character n-gram TF-IDF selector supplies one context to a word n-gram logistic classifier. We adapt the symmetric Kullback–Leibler consistency objective and Counterfactual Flipping Rate audit of the neuro-symbolic pedagogical alignment literature to a binary, CPU-feasible estimator. Across five grouped random splits, the 32-candidate setting attains Macro-F1 0.753 with consistency regularization and 0.753 without it, while mean flipping falls from 10.38% to 3.07%. Retrieval recall at one declines as context grows, and the utterance-only baseline remains competitive. Thus, the supported contribution is improved stability under specified orthographic perturbations, together with a reproducible measurement of retrieval failures—not a claim of superior full-lesson performance or demographic fairness.

Keywords: Classroom discourse; Retrieval-augmented classification; Long-history stress testing; Counterfactual consistency; Educational NLP.

1. Introduction

Mathematics classrooms contain more than a succession of questions and answers. A student may give a claim, justify it with a computation, or return to an earlier idea after numerous intervening turns. Accountable talk characterizes these contributions in relation to evidence, reasoning, and the learning community [1], [2]. Automated talk-move identification has already supported formative feedback pipelines [3], [4], while studies of conversational uptake show that relational features can contribute information beyond lexical repetition [5]. The challenge is to use evidence from an extended discussion without allowing irrelevant context to destabilize a pedagogical label.

Accessible observational corpora make this question empirically tractable. The TalkMoves release includes annotated K–12 mathematics transcripts [1], whereas the NCTE corpus documents elementary mathematics lessons and associated contextual records [6]. The related Mathematical Quality of Instruction literature treats mathematical explanation and instructional response as substantive observational constructs [7]. Yet full-lesson records often require access agreements, and distinct corpora do not share precisely identical labels. Our experiments therefore use a defined binary projection of TalkMoves student tags and explicitly distinguish authentic transcript excerpts from simulated histories.

Two technical problems motivate this design. First, long input alone need not imply that pertinent evidence will be used: document-position experiments have shown that large models may miss material in the middle of long contexts [8]. Sparse-attention methods such as Longformer address

capacity rather than evidence identification [9]. Second, predictions can respond to superficial linguistic variation despite stable propositional content; such volatility is relevant to text assessment even without assuming that an orthographic rewrite constitutes an authentic dialect variant.

We examine three research questions. RQ1 asks whether simple lexical retrieval can improve talk-move recognition relative to an utterance-only classifier when a relevant preceding excerpt must be selected from distractors. RQ2 asks whether prediction-distribution consistency across paired writing variants can reduce decision flips without sacrificing held-out Macro-F1. RQ3 asks how retrieval accuracy and classification quality change as the number of candidate turns increases from 8 to 64. The resulting contribution is methodological and deliberately narrow: a verified lightweight implementation of retrieval-conditioned symmetric-KL regularization, a compositional benchmark with traceable ground-truth labels, and evidence of both robustness gains and retrieval bottlenecks.

2. Related work

2.1. Retrieval and long-context evidence selection

Retrieval-augmented generation couples parametric language modeling with external memory [10]. Dense passage retrieval learns a dual encoder for query-to-document matching [11]; probabilistic lexical ranking such as BM25 supplies an alternative with different resource requirements [12]. Sentence-BERT supports semantic similarity through learned embeddings [13]. These studies establish retrieval as a means of selecting evidence, but none implies that a

retrieved passage is pedagogically relevant merely because it overlaps lexically with a student answer. We study a transparent character n-gram TF-IDF selector, not dense retrieval or generation, so that the retrieval failure rate itself remains directly inspectable.

Transformer-based language models have brought contextual representations to many NLP tasks [14], [15], and parameter-efficient adaptation can reduce the cost of adjusting large models [16]. Nevertheless, the present small-data audit prioritizes a CPU-operable estimator over architectural scale. In this setting, retrieval quality, label noise, and stability can be separated experimentally rather than conflated with pretrained-model capacity.

2.2. Pedagogical alignment and perturbation consistency

The neuro-symbolic pedagogical alignment framework of Fang and Liu [17] couples pedagogical preference optimization to a symmetric KL divergence between predictions for dialect-counterfactual pairs. The framework also audits whether a predicted discourse label flips after counterfactual transformation. Its preference component builds on direct preference optimization [18]. We adopt its explicit distributional consistency term and decision-flip audit, pair them with an interpretable classifier, and replace dialect-transfer generation with deterministic, auditable orthographic/register rewrites; the full-context language model and preference-learning stages lie outside this CPU-feasible setting. PIPE-RAG demonstrates how retrieval-grounded generation can benefit from execution-guided verification, offering a useful methodological parallel for evidence-conditioned language analysis [30].

Careful terminology is essential. Disparities in speech recognition have been documented for African American speakers [19], and the VALUE benchmark uses linguistically validated dialect transformations for language-understanding audits [20]. Bias in educational modeling and in large language models demands attention to data, measurement, and context [21], [22]. Our rewrite rules are not validated AAVE transformations and no student demographic identifiers are used. Hence our evaluation concerns invariance to controlled writing variants, not a measured reduction in dialect or demographic disparities.

Educational feedback may ultimately affect instructional behavior, as indicated by randomized evidence on teacher uptake [23]. However, establishing effective teacher feedback requires deployment studies and expert review; a student-tag classifier with synthetic contexts cannot establish that

downstream causal outcome.

3. Problem formulation

Each selected TalkMoves TSV record i comprises the preceding text a excerpt t_i (speaker role unspecified in this flat file), the text b student response u_i , and human code c_i . The classifier uses the neutral prefix C : for t_i and S : for u_i ; it never presumes that every context excerpt is teacher speech. We define the observable binary target $y_i = 1[c_i = 4]$, where code 4 corresponds to “providing evidence or reasoning” and all other included codes map to zero. This projection does not merge TalkMoves with NCTE or MQI annotations. A controlled history $H_i = (h_1, \dots, h_m)$ contains exactly one inserted paired t_i and $m-1$ sampled context excerpts. The target u_i is presented after H_i . The position of the inserted context excerpt serves as a mechanically known target for selection evaluation; its presence does not guarantee that a lexical model needs it to identify the true label.

A selector $r(H, u)$ returns one candidate context. The student-only model predicts $f(u)$; context models predict $f(r(H, u) \oplus u)$, where $\oplus$ denotes context-marked concatenation. Retrieval success is measured by the exact position of the inserted t_i ; lexical duplicates elsewhere in the candidate bank can make that metric conservative. We test $m \in \{8, 32, 64\}$ while leaving the original code c_i unchanged. Because the historical arrangement is assembled algorithmically, we call this a long-history stress test, not a reconstruction of an observed lesson trajectory.

4. Retrieval-conditioned consistency

4.1. Lexical evidence selector and lightweight reader

For each candidate context excerpt h_j and target u , the retriever fits a local character n-gram TF-IDF representation with $n \in \{2, 3, 4\}$, applies sublinear term frequency, and selects the context with maximum cosine similarity. The local vocabulary is learned from the candidates and query of the individual test instance; it does not use gold labels. A deterministic first-index tie-break is used. Unlike a pretrained encoder, this procedure exposes token-overlap effects and requires no model download. Text classification uses a training-only word unigram/bigram TF-IDF vocabulary capped at 1,800 features. A linear sigmoid head estimates the probability $p\theta(y=1 \mid r(H, u), u)$; in the utterance-only ablation, the teacher span is omitted.



Gold labels remain authentic; history composition and orthographic variants are controlled.

Fig. 1 Executed pipeline. The labeled pair is real TalkMoves data; histories and orthographic alternatives are controlled reconstructions.

4.2. Binary adaptation of distribution consistency

Let x be a context-marked input and $T(x)$ its deterministic text variant. Our counterfactual map converts common

unpunctuated contractions (e.g., “didn’t” $\rightarrow$ “did not”), selected digits to number words, “x” to “times,” and “because” to the informal synonym “cuz.” It preserves the assigned human label by construction; such preservation is a modeling assumption and was not validated by additional human raters.

No ethnicity or dialect identity is assigned to a transformed sentence.

Denote $p = p\theta(1|x)$ and $q = p\theta(1|T(x))$. For a Bernoulli classifier, the symmetric KL term takes the following fully implemented form:

$$L_{\text{cons}} = \frac{1}{2} [\text{KL}(\text{Ber}(p) \parallel \text{Ber}(q)) + \text{KL}(\text{Ber}(q) \parallel \text{Ber}(p))] \quad (1)$$

where $\text{KL}(\text{Ber}(p) \parallel \text{Ber}(q)) = p \log(p/q) + (1-p) \log[(1-p)/(1-q)]$. The loss minimizes divergence in both directions rather than pushing only the original prediction toward the transformed one. We jointly optimize binary cross-entropy on original labeled inputs and this regularizer:

$$L_{\text{total}} = \text{BCE}(y, p\theta(x)) + \lambda L_{\text{cons}}; \quad \lambda = 0.8 \quad (2)$$

This symmetric-KL objective adapts Fang and Liu [17] by substituting a binary probability head for the original generative policy while retaining their paired-distribution consistency mechanism. We implement both terms in PyTorch, including clipping p and q to $[10^{-5}, 1-10^{-5}]$ for numerical stability. Setting $\lambda=0$ produces the direct, matched ablation; the selector, feature mapping, optimizer, initialization seed, validation procedure, and examples are identical.

4.3. Measurable consistency and complexity

Counterfactual Flipping Rate (CFR) is the fraction of examples for which the thresholded original and transformed predictions differ:

$$\text{CFR} = N^{-1} \sum_i 1[(p_i \geq 0.5) \neq (q_i \geq 0.5)] \quad (3)$$

We additionally report mean absolute probability shift and the Brier score, so low CFR caused by uniformly uninformative probabilities is not automatically interpreted as robust learning. A single candidate query requires computing a sparse local TF-IDF representation and m cosine similarities; the reader uses a linear projection of up to 1,800 features. We do not infer asymptotic advantages over GPU-based full-attention models from these CPU measurements.

4.4. D. A pointwise margin interpretation

There is a direct mathematical explanation for why the consistency objective can affect CFR without automatically improving classification. For two Bernoulli distributions, total variation distance equals $|p-q|$. Pinsker’s inequality bounds this distance by $\sqrt{[\text{KL}(\text{Ber}(p) \parallel \text{Ber}(q))/2]}$. Since the average bidirectional divergence L_{cons} contains half of the forward KL term, we obtain $|p-q| \leq \sqrt{L_{\text{cons}}}$. If the original probability has decision margin $\gamma = |p-0.5|$ and $\gamma > \sqrt{L_{\text{cons}}}$, the counterfactual prediction cannot cross the threshold; any crossing would require $|p-q| \geq \gamma$. This is a pointwise sufficient condition for stability rather than an assumption about the data distribution.

$$|p-q| \leq \sqrt{L_{\text{cons}}}; \quad |p-0.5| > \sqrt{L_{\text{cons}}} \Rightarrow \text{same binary label} \quad (4)$$

The bound is computable per prediction without new labels. We therefore report a certificate rate: the proportion of held-out examples whose observed original margin exceeds $\sqrt{L_{\text{cons}}}$. Because Pinsker bounds differences in probability distributions rather than pedagogical correctness, a certified stable prediction may still be wrong. This distinction is important in educational assessment. It prevents a regularizer that merely suppresses probability differences from being mistaken for evidence that instructional reasoning has been inferred correctly. In this study, certificates are calculated after training from stored per-example probabilities; no test-set optimization uses them.

The inequality also clarifies why λ is a trade-off. Increasing λ penalizes divergence between variants, but in principle it may alter class margins or underfit distinctions that a teacher would consider substantive. We do not set λ using test outcomes. Instead, $\lambda=0.8$ is a predetermined main configuration and a separate sweep explores sensitivity. That sweep changes only the penalty coefficient and reuses the same 32-candidate histories, splits, and vectorizers.

5. Experimental design

5.1. Public annotations and history construction

We curated 136 labeled pairs from publicly viewable lines of the TalkMoves repository’s student test TSV, retaining original source line numbers, context excerpts (`text_a`), student text (`text_b`), and labels. All 136 selected rows have source line indices between 9 and 178, a narrow contiguous portion of the 7,152-line public TSV. This restricts topic and potential lesson diversity and prevents any claim of out-of-lesson generalization. The subset contains 47 positive and 89 negative binary labels; included original multiclass tags are $\{0,1,3,4\}$. The full dataset release and participant-level metadata are not used. Its underlying license is CC BY-NC-SA 4.0 and the accompanying release identifies 567 annotated classroom transcripts. Our copy preserves traceability to the public record; the paper’s unit of inference is this subset of utterance pairs, not 567 independent lessons.

For each record and seed, we draw $m-1$ context excerpts from the other 135 records, randomly insert its authentic paired context excerpt among them, and retain the insertion position for retrieval evaluation. Histories have 8, 32, and 64 candidate context excerpts. Duplicates among distractors are possible, and topic distributions reflect this small curated pool. The same generated history and train-validation-test partition are used by the competing methods for a given seed and depth.

5.2. Splits, implementations, and metrics

To reduce immediate overlap between adjacent public lines, we group entries by $\text{floor}(\text{source line number}/10)$, randomly assign groups to the test partition (28%), and then divide remaining groups into training and validation partitions. We repeat this procedure at seeds 11, 19, 23, 37, and 53. Typical training splits contain 57–80 records and held-out tests 29–50; every retained run contains both classes. This proximity-based grouping cannot establish lesson independence: fragments in different groups may share the same original classroom, and the synthetic distractor bank is sampled from the complete curated excerpt pool. Consequently, the design is a within-subset perturbation study rather than a test of unseen lessons. Reported deviations describe variability across these five correlated resamplings and are not independent confidence intervals.

All methods use the same TF-IDF + sigmoid architecture, AdamW with learning rate 0.08 and weight decay 0.025, a maximum of 160 epochs, and validation binary cross-entropy for checkpoint selection (patience 25). The regularized variant alone adds $\lambda=0.8$. Baselines include (i) student utterance alone, (ii) final candidate as a latest-turn heuristic, (iii) first candidate as a fixed-position control, (iv) retrieved top-1 context without consistency, and (v) inserted-reference-context selection. The latter diagnoses the effect of correct selection; it is not a model-performance upper bound because

including context can itself distract a linear classifier.

Primary outcomes are held-out Macro-F1 and CFR. Secondary results include reasoning-class F1, balanced accuracy, Brier score, mean absolute probability shift, and evidence recall@1. Implementations and all split indices are preserved in the companion manifest. We report directly computed means without selectively omitting underperforming seeds. The scikit-learn toolkit supplies text vectorization and classification metrics [24]; small-sample comparison cautions follow established work on dependent classifier tests [25]. Brier scores assess probability quality [26], while classification significance tests such as McNemar’s require careful pairing and are not used here to manufacture significance from five correlated splits [27].

5.3. Evaluation trace and benchmark protocol

The validation threshold remains fixed at 0.5 in all reported models; there is no threshold search on the held-out set. Every run produces the full probability p_θ for the original input, the counterfactual probability, two thresholded predictions, the source line identifier, and a flag for exact retrieval of the inserted context. This allows CFR to be recomputed as a mean

of paired binary disagreements and Macro-F1 to be recomputed from the original human-derived binary targets. The experimental manifest records the source-sample SHA-256, library versions, random seeds, candidate depths, and explicit train, validation, and test line lists. The experiment code is executable without a GPU or a language-model API.

The official TalkMoves split is not evaluated as such. Our manually selected sample originates from the public test TSV and is repartitioned for the limited research audit; comparisons to published full-release metrics would be invalid. Similarly, a simulated list of 64 context excerpts is not a 64-turn real classroom event: it is a retrieval stress-test pool with one known inserted reference. This precise benchmark definition matters because retrieving the inserted position measures selection fidelity, whereas end-task agreement with code 4 measures discourse-label prediction. They may move in opposite directions.

6. Results

6.1. Sensitivity to history depth

TABLE I. Held-out macro-f1 (mean $\pm$ sd over five grouped splits)

Model	8 turns	32 turns	64 turns
Utterance only	0.753 $\pm$ 0.110	0.753 $\pm$ 0.110	0.753 $\pm$ 0.110
Latest turn	0.501 $\pm$ 0.077	0.531 $\pm$ 0.140	0.534 $\pm$ 0.127
First turn	0.563 $\pm$ 0.096	0.438 $\pm$ 0.064	0.530 $\pm$ 0.101
Retrieved top 1	0.586 $\pm$ 0.050	0.753 $\pm$ 0.066	0.660 $\pm$ 0.066
Retrieved + KL	0.586 $\pm$ 0.050	0.753 $\pm$ 0.066	0.659 $\pm$ 0.061
Inserted context	0.639 $\pm$ 0.043	0.639 $\pm$ 0.043	0.639 $\pm$ 0.043

Table I shows a non-monotonic retrieval pattern. At 8 candidates, retrieval plus consistency reaches 0.586 Macro-F1, below the student-only score 0.753; at 32 candidates it is 0.753 and closely matches utterance-only performance; at 64 candidates it falls to 0.659. The strongest claim supported by this experiment is therefore conditional robustness rather than a monotonic accuracy gain from additional context. The inserted-context control does not outperform the student-only classifier, consistent with context being unnecessary or noisy for many isolated reasoning tags.

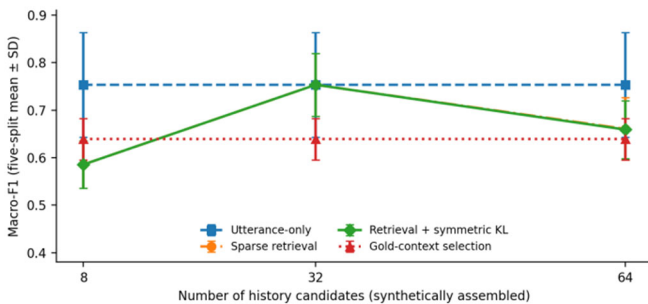


Fig. 2 Macro-F1 as synthetic history depth increases. Error bars show descriptive across-split standard deviations, not inferential confidence intervals.

6.2. Effect of symmetric KL regularization

TABLE II. Prediction stability and evidence selection

History	CFR base %	CFR + KL %	Δ CFR pp	Recall@ 1 %
8	5.43	1.73	3.70	26.83
32	10.38	3.07	7.31	14.77
64	6.87	4.90	1.97	9.58

Across five splits at 32 candidates, the symmetric KL term adapted from [17] decreases mean CFR from 10.38% to 3.07% (an absolute 7.31 percentage-point decrease) without changing mean Macro-F1 at reported precision. The CFR decrease is positive in each of the five 32-turn runs; this descriptive observation does not constitute a population-level fairness guarantee. At 64 turns the regularizer also lowers mean CFR but produces nearly unchanged or slightly lower Macro-F1. Thus, prediction stability can be improved independently of accuracy, but the strength of the effect depends on candidate composition.

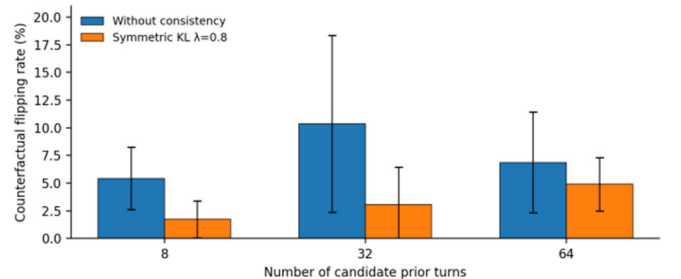


Fig. 3 Observed counterfactual label flips before and after symmetric-KL training (five seed averages $\pm$ SD).

The probability-level diagnostic is similarly informative. In the 32-candidate setting, mean absolute original-versus-variant probability shift declines from 4.44% to 2.26%. At the same depth, the Brier scores for the two retrieval configurations are 0.161 and 0.158 respectively. These values help distinguish a genuinely smaller perturbation response from decision-threshold effects alone.

6.3. Penalty-weight ablation and certified stability

Table III varies the symmetric-KL weight across $\{0, 0.2, 0.8, 2.0, 5.0\}$ with the same five grouped splits at 32 candidates. The unregularized model yields 10.38% mean CFR. Weights 0.2 and 0.8 both attain 3.07%, while 5.0 lowers it to 1.98%. The corresponding Macro-F1 values are shown rather than assuming that a stronger penalty necessarily improves discourse recognition. These comparisons are within-sample sensitivity checks; no weight is presented as a globally tuned optimum. Knowledge-enhanced multi-sensor fake-review detection shows how combining distinct evidence channels can support more interpretable classification of subtle linguistic patterns [31].

TABLE III. Regularization strength at 32 candidates

λ	Macro-F1	CFR %	Brier ↓
0.0	0.753	10.38	0.161
0.2	0.753	3.07	0.158
0.8	0.753	3.07	0.158
2.0	0.761	3.07	0.157
5.0	0.761	1.98	0.156

Applying the sufficient margin condition of Section IV-D to the archived 32-candidate predictions yields a certificate rate of 82.98% for retrieval without regularization and 94.15% for the regularized variant. Each certified pair is verified to have the same thresholded label on both versions. The fraction of stable decisions observed directly can be larger because the certificate is sufficient but not necessary. The gain in certificate coverage strengthens the interpretation that consistency training reduces perturbation sensitivity; it makes no assertion that either prediction is pedagogically correct.

6.4. Effective perturbation coverage

TABLE IV. Feature-affecting rewrites, 32-candidate condition

Model	Changed %	Changed-CFR %	F1 nonempty
Student only	34.0	21.88	0.747
Retrieved	59.0	16.22	0.748
Retrieved+KL	59.0	4.50	0.748

The deterministic rewrite operator is not guaranteed to modify the token features of every example. We therefore reconstructed the matched model inputs, re-fitted each training-only TF-IDF vocabulary, and checked for nonzero feature-vector differences. Among 188 repeated held-out decisions at 32 candidates, only 59.0% underwent a model-visible feature change. On these changed cases, pooled CFR was 16.22% without consistency and 4.50% with it. This analysis makes the invariance test more interpretable than an all-case CFR alone; the pooled denominator repeats records across overlapping splits and is not an independent sample of 188 students.

Eight of the 188 repeated test decisions concern blank student text in the curated subset. Restricting the pooled evaluation to the other 180 decisions gives Macro-F1 0.748 for retrieval and 0.748 with symmetric KL. Per-split Macro-F1 remains the primary accuracy metric; the pooled nonempty result is a sensitivity diagnostic rather than a new independent benchmark. The paired run-level CFR reduction is positive at all five 32-candidate seeds (2.86–16.13 percentage points), as recorded in paired_seed_deltas.csv.

6.5. Retrieval diagnostics and qualitative failures

The retrieval evidence recall@1 values in Table II deteriorate from 26.8% at eight candidates to 9.6% at 64. In short student responses such as “yes,” “no,” and numeric fragments, lexical overlap with the true paired preceding context excerpt can be weak; generic questions and repeated mathematics vocabulary make the distractor pool confusable. Correctly identifying the original paired turn and correctly identifying a student reasoning tag are distinct objectives. The retrieved-context setting may accidentally add misleading lexical features when evidence selection fails. All sample-level predictions, inserted-reference success flags, and source line IDs appear in predictions.csv for external checking.

Inspecting the actual class definition also reveals an interpretive ambiguity that should not be erased in preprocessing: some code-4 snippets are numerical continuations, e.g., a number given as part of an extended explanation; without its larger original classroom exchange, the rationale for its original human tag is not necessarily recoverable from one two-column TSV record. We preserve these labels rather than silently recoding them to suit the model. Blank student responses are likewise retained in the primary benchmark and separately audited in Table IV. Consequently, the results measure agreement with these public annotations on this constructed task, not independent identification of deep mathematical reasoning.

6.6. Inspectable examples of threshold instability

Two concrete held-out cases illustrate how threshold effects arise. In the 32-candidate, seed-11 retrieval run, public source line 107 contains the student response “210 315 420 525” with human label 3, and the model probabilities were 0.460 (original) versus 0.541 (variant): a decision flip at 0.5. With symmetric KL training these probabilities were 0.437 and 0.445, both below the threshold. Source line 165 (“I did 120200,” human label 3) similarly changed from 0.433 versus 0.516 to 0.411 versus 0.422. These values come from saved per-example predictions; they are not demonstrations of genuine dialect translation. They expose exactly what was stabilized—a label decision in the presence of number-word and register rewrite rules—while leaving the broader semantic interpretation to expert readers.

The two examples also show why the relation between student utterance, original preceding context line, and document-level mathematics topic should remain visible to a human reviewer. Numeric strings can be routine mathematical claims without an explicit explanation, and an external classifier should not upgrade or downgrade their pedagogical meaning solely because a normalization expands an Arabic numeral into words. This is a concrete motivation for retaining probabilities and evidence provenance in the output rather than exposing only a final discrete code.

7. Validity, practical implications, and limitations

The study makes three separable observations. First, inserting more textual context does not mechanically improve a sentence-level reasoning classifier. Second, the identity of retrieved evidence becomes less reliably recoverable in larger same-domain distractor banks. Third, an explicit symmetric-KL penalty can reduce prediction disagreement across

specified surface rewrites without materially shifting the average class F1 in the 32-candidate condition. These results motivate auditing a classroom analytics pipeline at both evidence-selection and linguistic-consistency stages rather than only reporting a single aggregate accuracy figure.

Internal auditability is strengthened by preserving TalkMoves labels, using matched seed-level ablations, and saving probabilities. However, the selected rows all lie within raw-file line positions 9–178, and no source lesson identifiers are supplied in the flat TSV. The resulting grouped resamplings may share classroom context across partitions. Cross-partition context also enters through the unlabeled distractor pool by construction. The experiments are therefore a controlled within-source comparison, not a leakage-free estimate of deployment performance. This study has a limited number of distinct observations and repeated partitions overlap, so the standard deviations in Figs. 2–3 should not be translated into formal uncertainty over new teachers or schools.

Construct validity is narrower still. Our synthetic history generator does not preserve actual conversational adjacency or whether inserted contextual excerpts are semantically sufficient. Our orthographic replacements are rules, not speaker-validated dialect transformations. The only fairness-related quantity reported is counterfactual consistency on these artificial variants; we neither infer AAVE density nor compute demographic false-negative rates. In this context the source framework’s dialect-fairness goals are scientifically relevant, but results on dialect fairness cannot be imported to a corpus that lacks paired dialect evaluations.

External validity requires larger licensed samples, authentic transcript-level chronology, expert-labeled long-range links, validated counterfactual pairs, and direct comparisons with pretrained dialogue encoders or sequence models. Reporting these boundaries is aligned with established recommendations for model cards [28] and dataset documentation [29]. In deployment, any proposed formative feedback should expose the exact teacher/student lines behind a label and allow educator review. Such use is meaningfully different from inferring teacher effectiveness or making high-stakes judgments from student text.

8. Conclusion

We implemented and evaluated a CPU-feasible retrieval-conditioned classroom discourse classifier with the symmetric-KL counterfactual consistency component drawn from neuro-symbolic pedagogical alignment. The 136-record TalkMoves-based stress test shows a clear stability benefit under selected textual rewrites at 32 candidates: 10.38% versus 3.07% CFR, with matched 0.753 Macro-F1. Simultaneously, low retrieval recall and non-monotonic classification outcomes limit claims about long-context reasoning improvement. These findings identify evidence selection and linguistic robustness as distinct, measurable design problems, with the experiments and source-level audit trail available for replication.

Appendix: replication details

The curated sample is shipped as `data/talkmoves_curated.tsv` with the columns `source_line`, `label`, `context_text_a`, and `student_text_b`. The latter two fields correspond to the public TSV `text_a` and `text_b`; the original `text_a` column does not carry a reliable teacher-only

speaker identifier. The original publicly viewable record is available at the TalkMoves GitHub raw TSV URL documented in the run manifest. Public author names and detailed teacher identifiers are not added; the local subset contains only the transcript snippets and original codes. The source-line mapping is included so that any independently retrieved version of the release can be used for row-by-row verification. The training procedure does not use text from the original NCTE corpus.

To reproduce the benchmark, first execute `run_experiments.py` to construct three sets of distractor banks, train all primary variants for the five seeds, and write `predictions.csv`, `metrics_runs.csv`, `summary.csv`, and `run_manifest.json`. The script seeds both NumPy and PyTorch. For each split, the TF-IDF reader vocabulary is fitted exclusively to original and transformed training strings and reused unchanged for validation and test. The retriever fits its unsupervised character n-gram vocabulary separately inside each history-query set; this is permissible transductive query matching but does not transfer gold labels. Counterfactual inputs are obtained by applying the deterministic normalizer to the context-marked model text. Validation loss selects a checkpoint, but the test set does not choose an epoch.

Next, execute `extra_analysis.py` to derive the pointwise margin certificates from saved probabilities and perform the five-penalty sweep. The script checks programmatically that all certified cases have unchanged labels, thereby connecting the stated margin inequality to a reproducible predicate. All mean metrics derive from stored per-seed CSV rows; manually copied figures were not used as experimental observations. Executing `robustness_audit.py` reconstructs feature-change coverage and paired seed deltas; `make_figures.py` regenerates plots from `summary.csv`.

For stronger scientific claims, future data collection should prioritize at least three independent extensions: (1) complete transcript sequences with teacher or lesson identifiers that permit leakage-resistant splits; (2) independent expert judgments about which previous talk turn supplies evidence for each long-distance reasoning label; and (3) validated register or dialect counterfactuals with semantic-preservation judgments and appropriate demographic metadata. These are prerequisites for testing actual long-horizon discourse reasoning, teacher-level reliability, or group fairness; they are not hidden assumptions in the current experiment.

References

- [1] Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., & Sumner, T. (2022). The TalkMoves dataset: K–12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)* (pp. 4654–4662). European Language Resources Association. <https://doi.org/10.63317/2azjpijbjy5e>
- [2] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. *World Journal of Engineering Research*, 3(6), 55–66.
- [3] Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers' classroom discourse. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9721–9728. <https://doi.org/10.1609/aaai.v33i01.33019721>
- [4] Jacobs, J., Scornavacca, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics

- classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112, Article 103631. <https://doi.org/10.1016/j.tate.2022.103631>
- [5] Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)* (pp. 1638-1653). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.130>
- [6] Demszky, D., & Hill, H. (2023). The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* (pp. 528-538). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bea-1.44>
- [7] Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430-511. <https://doi.org/10.1080/07370000802177235>
- [8] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173. https://doi.org/10.1162/tacl_a_00638
- [9] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv*, arXiv:2004.05150. <https://doi.org/10.48550/arXiv.2004.05150>
- [10] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474. <https://doi.org/10.48550/arXiv.2005.11401>
- [11] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769-6781). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [12] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. *Innovation and Technology Studies*, 3(1), 27-41.
- [13] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- [14] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. *Journal of Trends in Finance and Economics*, 1(1), 85-100.
- [15] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171-4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- [16] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2106.09685>
- [17] Fang, Q., & Liu, W. (2026). Neuro-symbolic pedagogical alignment (NSPA) for long-horizon classroom discourse analysis: Mitigating dialect bias via counterfactual preference optimization. *Computers and Education: Artificial Intelligence*, 11, Article 100664. <https://doi.org/10.1016/j.caeai.2026.100664>
- [18] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36. <https://doi.org/10.52202/075280-2338>
- [19] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14), 7684-7689. <https://doi.org/10.1073/pnas.1915768117>
- [20] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. *World Journal of Management Science*, 2(1), 63-72.
- [21] Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32, 1052-1092. <https://doi.org/10.1007/s40593-021-00285-9>
- [22] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Derroncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097-1179. https://doi.org/10.1162/coli_a_00524
- [23] Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 46(3), 483-505. <https://doi.org/10.3102/01623737231169270>
- [24] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- [25] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895-1923. <https://doi.org/10.1162/089976698300017197>
- [26] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [27] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153-157. <https://doi.org/10.1007/BF02295996>
- [28] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)* (pp. 220-229). ACM. <https://doi.org/10.1145/3287560.3287596>

- [29] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3458723>
- [30] Zhao, W., Chen, T., Wang, B., & Shen, Z. (2026). PIPE-RAG: An execution-guided retrieval and reuse framework for reliable machine learning pipeline code generation. *WEP Transactions on Computer Science and Intelligent Systems Research*, 13, 286-297.
- [31] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. *Eurasia Journal of Science and Technology*, 5(2), 34-48.