

Redact or Keep? A Fully Local AI Cascade for Educational Dialogue De-Identification

Haocheng Zhang¹, Zhuqian Zhou¹, Kirk Vanacore¹, Bakhtawar Ahtisham¹ and René F. Kizilcec¹

¹Cornell University, Ithaca, NY, USA

Abstract

Educational dialogue is a valuable but sensitive resource for research, yet de-identification is difficult because personally identifiable information often overlaps with curricular content: “Riemann” may refer to a student or to a mathematical concept. Commercial LLMs can handle this ambiguity but require data egress, while local NER systems preserve governance but often over-redact educational content. We propose a fully local cascade that separates de-identification into high-recall candidate proposal and contextual Redact/Keep review. The first stage combines two lightweight encoders with rule-based patterns to over-generate candidate spans; the second stage decides whether each span is real PII using surrounding dialogue and speaker role. Across math tutoring transcripts from two platforms, the strongest local configuration achieves a macro F1 of 0.958, compared with 0.767 for a same-family LLM-only baseline and 0.706 for a commercial API. On a targeted challenge set for curricular-personal name ambiguity, it degrades by only 0.03 F1 versus 0.19–0.25 for smaller reviewers. These results show that for educational de-identification, problem formulation can matter more than model scale, enabling privacy-preserving deployment without external APIs.

Keywords

de-identification, PII detection, educational NLP, local deployment, name disambiguation

1. Introduction

Authentic educational dialogue is a rich but sensitive resource for research. The same transcripts that capture how students learn often capture what they disclose about themselves, requiring removal of personally identifiable information (PII) before release. In educational dialogue, de-identification is harder than standard named-entity recognition because the same surface form may refer either to curricular content that should be preserved (e.g., “Riemann” in “Riemann sums”) or to a real student that must be redacted (“Hi, I’m Maria Riemann”). A practical system must therefore balance four requirements: detection accuracy, curricular preservation, privacy governance, and deployment cost.

Commercial LLM APIs improve detection performance [1, 2, 3], but sending student transcripts to third-party providers introduces data governance and compliance risks that many institutions worry about. Generic local named entity recognition (NER) systems avoid data egress but often fail on curricular-personal ambiguity. Existing approaches, therefore, trade governance for accuracy or vice versa.

To address these constraints, we propose a fully local cascade framework for educational dialogue de-identification. A *union* proposer combines two lightweight encoders (DeBERTa and ModernBERT) with direct rules to intentionally over-generate candidate PII spans, optimizing for recall at the expense of precision (recall-first). A second-stage *cascade-aligned reviewer* uses an LLM to inspect each candidate in context and emits a binary REDACT/KEEP decision. This formulation prioritizes recall-first but allows the reviewer to prevent over-reductions when necessary by converting open-ended entity recognition into constrained privacy triage. We compare three reviewer configurations—an encoder reviewer, a 4B LLM

AI for Education Day, SIGKDD 2026, August 12, 2026, Jeju, Korea

✉ hz782@cornell.edu (H. Zhang); zz968@cornell.edu (Z. Zhou); kpv27@cornell.edu (K. Vanacore); ba453@cornell.edu (B. Ahtisham); kizilcec@cornell.edu (R. F. Kizilcec)

📄 0009-0009-8580-1571 (H. Zhang); 0000-0002-8045-6213 (Z. Zhou); 0000-0003-0673-5721 (K. Vanacore); 0009-0007-9331-5069 (B. Ahtisham); 0000-0001-6283-5546 (R. F. Kizilcec)



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

reviewer trained with Low-Rank Adaptation (LoRA), and a 31B LoRA-trained LLM reviewer—against same-family LLM-only baselines and a commercial API baseline.

This setting is a data-mining problem because the central challenge is not only detecting entities but also preserving analytic utility in large-scale educational dialogue while satisfying privacy constraints. The paper addresses the following research questions (RQs):

- **RQ1 — Cascade vs. LLM-Only.** Holding the base model family constant, does a cascaded proposer–reviewer framework outperform single-step LLM-only full-dialogue extraction?
- **RQ2 — Robustness on ambiguous names.** On a pre-registered challenge set where real student names overlap with curricular-content-based names, which reviewers remain robust?
- **RQ3 — Deployment footprint.** Under commodity hardware constraints, what are training and inference costs, and can the full pipeline remain fully on-device?

2. Related Work

2.1. From Mature De-identification Domains to Educational Dialogue

De-identification is well established in healthcare, where shared tasks such as i2b2/UTHealth have supported mature sequence-labeling approaches [4]. However, clinical assumptions do not transfer cleanly to educational dialogue: the same surface form may refer to a student, a public figure, or a fictional character in a word problem [5]. Educational de-identification is therefore less a pure span-detection task than a contextual privacy decision.

Educational de-identification often has substantially different affordances that require distinct approaches. PIIvot [6] addresses tutoring-dialogue anonymization through recall-first potential-PII labeling and surrogate replacement, but does not make the explicit REDACT/KEEP preservation decision required for curricular-content disambiguation. MathEd-PII [3] targets numeric ambiguity in mathematical expressions, rather than name disambiguation. Ji et al. [2] and Singhal et al. [1] evaluate GPT-4o-mini and GPT-4 on educational corpora [7], achieving high accuracy but sending transcripts through commercial APIs, trading governance for accuracy. However, none of these evaluate the curricular-personal name ambiguity typical of educational contexts, nor do they report an end-to-end local deployment.

2.2. From Entity Detection to Privacy Triage

Off-the-shelf detectors like spaCy [8] and Microsoft Presidio [9] are local and fast but trained on news-like text. Their default behavior tends to over-redact curricular terms while providing no explicit keep-vs-redact distinction for ambiguous mentions. Customized Presidio recognizers and transformer-backed spaCy may improve lexical recall, but they still treat the task as entity detection rather than contextual privacy triage. In educational dialogue, identifying a candidate mention is often straightforward; deciding whether it should actually be redacted is the harder problem.

Two-stage NER systems provide a useful structural template, separating span proposal from subsequent classification or typing. Prior variants include cascaded fine-grained NER, T2-NER, TadNER, and ToNER [10, 11, 12, 13]. In these systems, the second stage typically performs entity typing. Our work instead uses the second stage for privacy verification: deciding whether a candidate span should be redacted or preserved in context.

3. Proposed Pipeline

3.1. Pipeline Overview

As noted above, we formulate de-identification as a two-stage classification problem over dialogue turns. Figure 1 illustrates the pipeline. Each dialogue turn is processed in parallel by the encoder proposer (a union of fine-tuned DeBERTa and ModernBERT; Stage 1A) and a set of direct RegEx rules (Stage 1B). The candidate pool is the union of spans from either source, intentionally favoring high

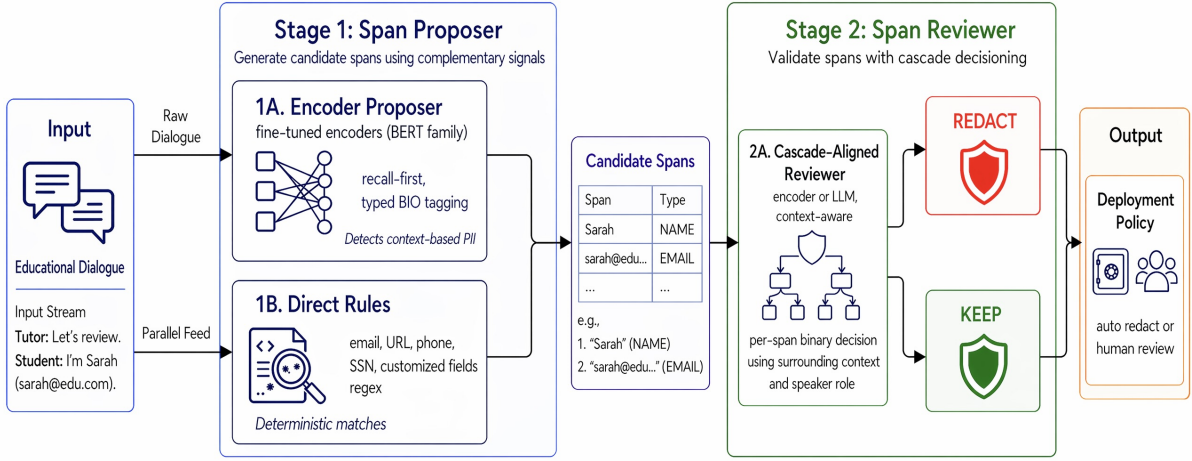


Figure 1: End-to-end cascade pipeline. A recall-first *union* encoder proposer (DeBERTa-v3-base + ModernBERT-base; stage 1A) plus direct rules (stage 1B) emit candidate spans; a cascade-aligned reviewer (stage 2A) decides REDACT/KEEP for each candidate using surrounding context and speaker role; a deployment-time policy applies the final privacy action.

recall over early precision. Each candidate span is then passed to the reviewer (Stage 2) together with its surrounding context, speaker role, and turn metadata, while rule-matched spans are high-precision structured identifiers and are emitted as REDACT directly, bypassing the reviewer (Appendix A). Rather than performing open-ended entity extraction, the reviewer makes a constrained binary REDACT or KEEP decision. This reframes the task from entity recognition to privacy triage. Spans predicted as PII are labeled REDACT and then handled according to a deployment-time policy. In an *auto-redact* setting, all such spans are automatically redacted to minimize disclosure risk. In a *human-review* setting, low-confidence cases are routed to manual verification, allowing institutions to balance privacy protection against curricular preservation. As constructed, the entire pipeline can be run locally with no external API calls.

3.2. Stage 1: Span Proposer

The proposer is a recall-first union of two fine-tuned encoders, DeBERTa-v3-base [14] and ModernBERT-base [15], both trained with typed BIO token classification, a sequence-labeling schema that marks each token as the Beginning, Inside, or Outside of a typed entity span (e.g., B-NAME, B-ADDRESS). Any span proposed by either encoder is passed forward, intentionally favoring recall over early precision.

In parallel, a deterministic rule-based proposer handles structured identifier types with strong lexical patterns, including EMAIL, URL, IP_ADDRESS, PHONE_NUMBER, IDENTIFYING_NUMBER, and customized lexicon-based types (e.g., in our example, we also redacted the name of the TUTOR_PROVIDERS that generated the data). These types are sparse in educational data and well-suited to high-precision pattern matching, making rules more reliable and efficient than learned detection.

3.3. Stage 2: Cascade-Aligned Reviewer

3.3.1. Label Alignment

The reviewer operates on proposer-generated candidate spans rather than full dialogues. For each proposed span, overlap with a gold annotation (manually labeled PII spans; see Section 4.1) is mapped to REDACT (true positive), while no overlap is mapped to KEEP (false positive). This mechanically aligns training labels with the actual deployment decision: whether the span should be removed. This formulation converts open-ended NER into a constrained binary classification task. Instead of asking the model to extract all sensitive spans from raw dialogue, the reviewer only determines whether a proposed candidate is true PII. This reduces task complexity, improves alignment with privacy decisions,

and makes small local models substantially more practical. Reviewer training uses binary labels with focal loss [16] to address class imbalance.

3.3.2. Reviewer Configurations

We evaluate three reviewer configurations on the same union-proposer output: a RoBERTa reviewer (125M, full fine-tune, 3 seeds) [17], a Gemma E4B-it [18] reviewer (4B, LoRA [19, 20], 1 seed), and a Gemma-4-31B-it [18] reviewer (31B, LoRA, 3 seeds). Hyperparameters and the mlx-lm gradient-clipping patch required for 3-seed 31B training are in Appendix C. Both the 4B and 31B LoRA LLM reviewers share a context-aware system prompt (Appendix E) that instructs the model to preserve names used in educational content while still redacting real student identities.

4. Method

4.1. Data

4.1.1. Data Sources

We evaluate the pipeline on math tutoring transcripts from two large online tutoring providers with different instructional settings. Table 1 presents the platforms and their respective data sets. Platform A contains short, question-focused K–12 tutoring dialogues, while Platform B contains longer one-to-one K–12 tutoring dialogues from scheduled online lessons with extended multi-turn interaction. As a result, Platform B sessions are substantially longer and more conversational than Platform A. All transcripts were manually annotated for PII by three annotators using a shared codebook with disagreements resolved through consensus discussion.

Table 1

Dataset summary. “Candidate turns” are turns scored by the Stage-1 Span Proposer as potential turns with PII which are used as the input for the Stage-2 Span Reviewer. “Dialogues without PII” are retained as negative examples.

Split	Sources	#Dialogues	#Tokens	#Candidate Turns	#Turns with any PII	%Turns with NAME	#Dialogues without PII
Train	Platform A	800	950,886	92,735	1,295	83.7	314
	Platform B	305	1,654,347	152,430	5,379	99.4	10
Val	Platform A	100	134,006	12,980	242	75.6	43
	Platform B	76	408,124	38,622	1,130	98.7	3
Test	Platform A	100	111,313	11,606	165	81.2	39
	Platform B	100	513,275	47,578	1,445	99.4	7
Challenge Test	Platform A	50	100,571	9,017	117	99.1	7
	Platform B	80	490,421	39,005	2,743	100.0	0
Total		1,611	4,362,943	403,973	12,516	97.1	423

4.1.2. Canonical Evaluation Set

The main evaluation set contains 200 held-out dialogues: 100 from Platform A and 100 from Platform B. We retain dialogues with zero gold PII spans (46 total: 39 from Platform A and 7 from Platform B) as negative examples, since a practical de-identification system must correctly decide when no redaction is necessary. These cases are important for measuring over-redaction and curricular preservation rather than recall alone.

The evaluation is strongly dominated by NAME entities. On the canonical test set, 81.2% of gold spans in Platform A and 99.4% in Platform B are annotated as NAME. This reflects the real operational distribution of educational dialogue, where student names, tutor names, and family references are the most common privacy-sensitive mentions.

4.1.3. Challenge Set of Ambiguous Names

A central failure mode in educational dialogues is ambiguity between real student identities and curricular content-based names often used in mathematical word problems. This distinction cannot be resolved reliably through lexical cues alone. To isolate this failure mode, we construct a targeted challenge set of 130 dialogues (50 from Platform A and 80 from Platform B) in which real student names and curricular-content-based names explicitly co-occur. The set was assembled by manually screening approximately 300 candidate dialogues for this co-occurrence pattern. These dialogues were annotated using the same guidelines and annotator process as the canonical test set, but they are intended as a stress test rather than a representative deployment sample. Challenge-set dialogues were excluded from all training splits, and no hyperparameter decisions were made using challenge-set results. Table 1 summarizes the dataset composition.

4.2. Experimental Design

4.2.1. Baseline Conditions: LLM-Only Full-Dialogue Detectors

To establish baselines, we evaluated four LLM-only full-dialogue detectors, all using the same candidate-free span-extraction formulation. Each model received the entire tutoring dialogue as a single input example and directly produced a JSON-formatted `redact_spans` output. The four baselines included: (1) Gemma E4B-it (4B), trained with LoRA using data specified in Table 1; (2) Gemma-4-31B-it zero-shot prompting; (3) Gemma-4-31B-it with LoRA fine-tuning for one epoch using the same data specified in Table 1; and (4) Gemini 3.1 Pro Preview [21], evaluated as a commercial API baseline in zero-shot mode through a secure institutional LiteLLM gateway under a formal data use agreement (DUA). The three Gemma variants ran on local hardware and therefore satisfied the privacy and governance requirements of local deployment. The Gemini comparison was included only as a controlled external reference point under DUA-authorized data egress, rather than as a deployable baseline.

4.2.2. Experimental Conditions: Two-Stage Cascaded Framework

For RQ1, the base model was held constant — Gemma E4B-it and Gemma-4-31B-it appear in both cascade and LLM-only conditions — to isolate formulation from model scale. These formulations were evaluated on both the conical and challenge datasets (RQ2). For RQ3 (deployment footprint), we measured wall-clock training time, inference throughput, and resident memory on a single laptop. RoBERTa and 31B LoRA reviewers were trained with three random seeds (13, 42, 101); 4B LoRA and 31B LLM-only baselines used a single seed (42). The 4B reviewer was trained with a single seed due to compute budget; we treat it as a point estimate. For multi-seed configurations, all values in Tables 2 and 3 are means across the three seeds; per-seed standard deviations are in Appendix D, and hyperparameter details are in Appendix C.

4.3. Evaluation

To evaluate the detection performance, we report span-level precision, recall, and F1. A proposed span counts as a true positive if its token span overlap with any gold PII span; we use overlap matching rather than exact match to avoid penalizing benign boundary variation such as titles, possessives, and tokenization differences (e.g., when gold contains 'Mr. Smith' and a model identifies 'Smith'). This criterion is applied consistently across all models and experimental settings.

5. Results

5.1. RQ1: Performance of Cascade vs. LLM-Only Configurations

Table 2 shows that the cascaded proposer-reviewer formulation consistently outperformed LLM-only full-dialogue extraction when the base model family was held constant. This supports the central claim

that in tutoring dialogue de-identification, problem formulation matters more than model scale alone.

Table 2

Performance on the canonical evaluation test. “Union” = DeBERTa-v3-base + ModernBERT-base proposer candidates with direct RegEx rules (see Section 3.2 for details).

Configuration	Platform A			Platform B			Macro F1
	Precision	Recall	F1	Precision	Recall	F1	
<i>Proposer-Only</i>							
DeBERTa	.867	.946	.904	.959	.977	.968	.936
Union (DeBERTa & ModernBERT)	.714	.970	.823	.878	.981	.927	.875
<i>Proposer + Reviewer</i>							
Union + RoBERTa	.841	.960	.896	.953	.977	.965	.931
Union + Gemma E4B	.859	.994	.921	.956	.972	.964	.943
Union + Gemma 31B	.926	.978	.951	.952	.977	.965	.958
<i>LLM-Only</i>							
Gemma E4B (0-shot)	.880	.400	.550	.950	.120	.213	.381
Gemma 31B (0-shot)	1.000	.564	.721	.993	.093	.170	.446
Gemma 31B LoRA	.913	.891	.902	.952	.474	.632	.767
Gemini 3.1 Pro	.962	.770	.855	.984	.389	.557	.706
<i>Industry-Standard</i>							
Presidio (default)	.153	.824	.258	.299	.906	.450	.354
spaCy (en_core_web_sm)	.031	.770	.060	.051	.911	.097	.078

The union proposer behaves as intended: it improves recall relative to DeBERTa alone while reducing precision, shifting precision recovery to the reviewer stage.

Adding a reviewer recovered most of this precision loss while preserving high recall. The best overall system was *Union + Gemma 31B*, achieving the highest macro F1 of 0.958 (0.951 on Platform A, 0.965 on Platform B). The 4B reviewer also performed strongly (0.943 macro F1), showing that strong local performance does not require the largest model.

In contrast, LLM-only baselines showed severe recall collapse, especially on Platform B. The 4B LLM-only baseline reached only 0.381 macro F1, and even Gemma 31B LoRA reached 0.767 — still 0.191 below the 31B cascade reviewer. Gemini 3.1 Pro improved to 0.706 but remained below all cascade systems while requiring external data egress. This collapse correlates with dialogue length: per-dialogue recall for the 4B LLM-only baseline drops from 0.54 on dialogues under 2K characters to 0.11 on dialogues over 20K characters (Appendix B).

Industry-standard local baselines (Presidio and spaCy) achieved high recall but extremely low precision, heavily over-redacting curricular content. Overall, a recall-first proposer followed by binary REDACT/KEEP review proved substantially more effective than direct full-dialogue extraction.

5.2. RQ2: Performance on The Challenge Set of Ambiguous Names

Table 3 evaluates the failure mode typical to educational context: ambiguous names that may refer either to real student identities (REDACT) or to curricular-content-based names (KEEP).

The strongest system was again *Union + Gemma 31B*, which achieved the best macro F1 of 0.932. On Platform A, it improved F1 from 0.574 (DeBERTa proposer only) to 0.894, mainly by recovering precision: proposer-only precision was only 0.408 due to heavy over-redaction, while the 31B reviewer raised it to 0.899 while maintaining strong recall (0.892). On Platform B, it also achieved the best result (0.969 F1), showing that contextual review is essential for resolving ambiguous names.

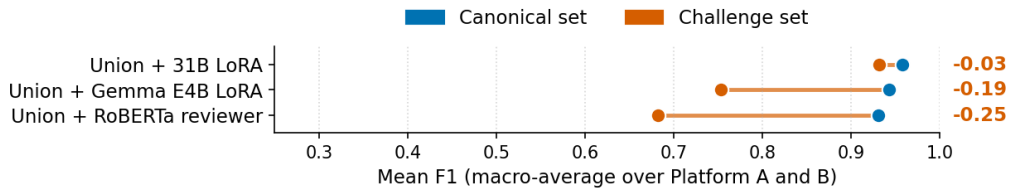
Smaller reviewers were much less robust. *Union + RoBERTa* and *Union + Gemma E4B* remained near 0.58 F1 on Platform A, showing little improvement over proposer-only detection. RoBERTa also reduced recall on Platform B (0.654), indicating that encoder reviewers often compounded proposer errors rather than correcting them.

Table 3

Performance on the challenge set of ambiguous names.

Configuration	Platform A			Platform B			Macro F1
	Precision	Recall	F1	Precision	Recall	F1	
<i>Proposer-Only</i>							
DeBERTa-only	.408	.966	.574	.985	.871	.925	.750
<i>Proposer + Reviewer</i>							
Union + RoBERTa	.425	.912	.580	.975	.654	.784	.682
Union + Gemma E4B	.410	.974	.577	.872	.991	.928	.753
Union + Gemma 31B	.899	.892	.894	.994	.945	.969	.932
<i>LLM-Only</i>							
Gemma E4B	.457	.319	.376	.960	.157	.270	.323
Gemini 3.1 Pro	.644	1.000	.784	.988	.939	.963	.873
<i>Industry-Standard</i>							
Presidio (default)	.099	.761	.175	.360	.883	.512	.348
spaCy (en_core_web_sm)	.022	.761	.044	.080	.883	.147	.096

The LLM-only 4B baseline performed worst overall (0.323 macro F1) because of severe recall collapse. The commercial API baseline (*Gemini 3.1 Pro*) performed strongly (0.873 macro F1), but still underperformed the fully local 31B cascade, especially on Platform A (0.784 vs. 0.894 F1).

**Figure 2:** Canonical-to-challenge F1 shift per configuration.

Canonical-to-challenge degradation (Figure 2) further highlights the robustness of *Union + Gemma 31B*: it drops by only .03 F1, compared with 0.19 for Gemma E4B LoRA and 0.25 for RoBERTa. It is the only configuration that maintains near-canonical performance, breaking the apparent near-equivalence among reviewers on the canonical test and establishing the 31B reviewer as the most robust reviewer by a substantial margin.

These results show that ambiguous-name resolution is not primarily a span-detection problem. Strong performance requires contextual REDACT/KEEP review rather than better candidate extraction alone.

5.3. RQ3: Deployment Footprint

To assess deployment feasibility, we measured wall-clock training time, inference throughput, and resident memory for each configuration on a single Apple M5 Max with 128 GB unified memory. Table 4 summarizes these costs alongside macro F1 for reference.

Table 4 shows that the best system remains feasible on a single Apple M5 Max: Union + Gemma 31B reaches 0.958 F1 with approximately 4.3 hours of training per seed and 18 GB inference memory. The 4B reviewer offers a strong efficiency tradeoff, retaining most of the accuracy at roughly half the memory and higher throughput. LLM-only models are less efficient in accuracy-per-resource terms: Gemma 31B LoRA uses similar memory to the 31B cascade but reaches only 0.767 F1. Overall, deployment cost is shaped less by model size alone than by whether model capacity is applied to localized Redact/Keep decisions.

Table 4

Training and inference cost on Apple M5 Max (128 GB, warm-start). Training time is wall-clock per seed. 31B cascade dial/s is reviewer-stage timing on the 200-dialogue canonical workload (proposer + rules add <5 s, negligibly).

Configuration	Training Time (h)	Throughput (#Dialogues/s)	Inference RSS (GB)	F1
<i>Proposer-Only</i>				
DeBERTa-only	~2.1	0.52	4.7	.936
<i>Proposer + Reviewer</i>				
Union + RoBERTa	~1.3	0.50	6.1	.931
Union + Gemma E4B	~2.8	0.36	9.5	.943
Union + Gemma 31B	~4.3	0.09	~18	.958
<i>LLM-Only</i>				
Gemma E4B	1.2	0.76	5.5	.381
Gemma 31B LoRA	1.8	0.06	~18	.767

6. Conclusion

We introduced a fully local cascade framework for de-identifying educational dialogue. The framework separates the task into high-recall candidate proposal followed by contextual Redact/Keep review. This formulation addresses a central challenge in educational data mining: preserving the analytic utility of large-scale learning data while satisfying privacy and governance constraints.

Across math tutoring transcripts from two large platforms, the cascade substantially outperformed same-family LLM-only extraction and a commercial API baseline. The strongest configuration, Union + Gemma 31B, reached 0.958 macro F1 on the canonical test set, compared with 0.767 for the same model used as a single-pass detector and 0.706 for the commercial API baseline. On a targeted challenge set for curricular–personal name ambiguity, the same configuration degraded by only 0.03 F1, while smaller reviewers degraded by 0.19 to 0.25 F1. These results suggest that, in this setting, problem formulation can matter more than model scale alone: strong de-identification requires not only finding candidate spans, but deciding whether those spans represent real privacy risk in context.

The deployment results further show that privacy-preserving de-identification need not require data egress to third-party APIs. The strongest configuration trained and ran on a single laptop-class machine, requiring approximately 4.3 hours of training per seed and 18 GB of inference memory. This makes the approach practical for institutions that need to process sensitive educational dialogue under local governance constraints.

Several limitations remain: First, the evaluation is limited to English-language math tutoring transcripts from two platforms; performance may differ in other subjects, languages, age groups, or instructional settings. Second, the cascade adds inference latency relative to a single-pass detector, which may matter for real-time applications even if it is acceptable for batch de-identification. Third, deployment-cost estimates are specific to Apple Silicon with unified memory and may vary on discrete-GPU infrastructure. Finally, the challenge set is intentionally targeted and relatively small, so its results should be interpreted as a stress test rather than a representative deployment estimate.

Future work should evaluate the framework across broader educational domains, expand challenge sets for additional ambiguity types, and study human-in-the-loop review policies for low-confidence cases. More broadly, this work shows that educational de-identification is not simply a generic NER problem; it is a contextual privacy-triage problem in which the goal is to protect learners while preserving the educational content needed for research and improvement.

Acknowledgments

This material is based upon initial work completed under National Science Foundation Grant No. 2321499, and support from the Gates Foundation and the Chan Zuckerberg Initiative. Any opinions, findings,

and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) and Claude (Anthropic) in order to: Grammar and spelling check, Paraphrase and reword, and Improve writing style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] S. Singhal, A. F. Zambrano, M. Pankiewicz, X. Liu, C. Porter, R. S. Baker, De-identifying student personally identifying information with GPT-4, in: Proceedings of the 17th International Conference on Educational Data Mining, International Educational Data Mining Society, Atlanta, Georgia, USA, 2024, pp. 559–565. URL: <https://educationaldatamining.org/edm2024/proceedings/2024.EDM-short-papers.57/>. doi:10.5281/zenodo.12729884.
- [2] Z. Ji, Y. Shen, K. R. Koedinger, J. Lin, Enhancing the de-identification of personally identifiable information in educational data, *Journal of Educational Data Mining* (2025). URL: <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/936>. doi:10.5281/zenodo.17114271.
- [3] Z. Zhou, K. Vanacore, B. Ahtisham, J. Lee, D. Pietrzak, D. Hedley, J. Dias, C. Shaw, R. Schäfer, R. F. Kizilcec, Utility-preserving de-identification for math tutoring: Investigating numeric ambiguity in the MathEd-PII benchmark dataset, *arXiv preprint arXiv:2602.16571* (2026). URL: <https://arxiv.org/abs/2602.16571>.
- [4] A. Stubbs, C. Kotfila, Ö. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1, *Journal of Biomedical Informatics* 58 (2015) S11–S19. URL: <https://pubmed.ncbi.nlm.nih.gov/26225918/>. doi:10.1016/j.jbi.2015.06.007.
- [5] A. Kovačević, B. Bašaragin, N. Milošević, G. Nenadić, De-identification of clinical free text using natural language processing: A systematic review of current approaches, *Artificial Intelligence in Medicine* 151 (2024) 102845. URL: <https://doi.org/10.1016/j.artmed.2024.102845>. doi:10.1016/j.artmed.2024.102845.
- [6] M. Zent, D. Smith, S. Woodhead, PIIvot: A lightweight NLP anonymization framework for question-anchored tutoring dialogues, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 27479–27488. URL: <https://aclanthology.org/2025.emnlp-main.1397/>. doi:10.18653/v1/2025.emnlp-main.1397.
- [7] L. Holmes, J. Wang, S. Crossley, W. Zhang, The cleaned repository of annotated personally identifiable information, in: Proceedings of the 17th International Conference on Educational Data Mining, 2024. URL: <https://educationaldatamining.org/edm2024/proceedings/2024.EDM-posters.88/index.html>.
- [8] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, *spacy: Industrial-strength natural language processing in python*, Software, 2020. URL: <https://spacy.io>.
- [9] Microsoft, Presidio: Data protection and de-identification sdk, Software documentation, 2024. URL: <https://microsoft.github.io/presidio/>.
- [10] P. Awasthy, T. Moon, N. Jian, R. Florian, Cascaded models for better fine-grained named entity recognition, *arXiv preprint arXiv:2009.07317* (2020). URL: <https://arxiv.org/abs/2009.07317>.
- [11] P. Huang, X. Zhao, M. Hu, Z. Tan, W. Xiao, T²-NER: A two-stage span-based framework for unified named entity recognition with templates, *Transactions of the Association for Computational Linguistics* 11 (2023) 1265–1282. URL: <https://aclanthology.org/2023.tacl-1.72>. doi:10.1162/tacl_a_00602.

- [12] Y. Li, Y. Yu, T. Qian, Type-aware decomposed framework for few-shot named entity recognition, arXiv preprint arXiv:2302.06397 (2023). URL: <https://arxiv.org/abs/2302.06397>.
- [13] G. Jiang, Z. Luo, Y. Shi, D. Wang, J. Liang, D. Yang, ToNER: Type-oriented named entity recognition with generative language model, arXiv preprint arXiv:2404.09145 (2024). URL: <https://arxiv.org/abs/2404.09145>.
- [14] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, 2023. URL: <https://openreview.net/forum?id=sE7-XhLxHA>.
- [15] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663 (2024). URL: <https://arxiv.org/abs/2412.13663>.
- [16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988. URL: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html. doi:10.1109/ICCV.2017.324.
- [17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: <https://arxiv.org/abs/1907.11692>.
- [18] Gemma Team, Google DeepMind, Gemma 4: Open models for on-device and local deployment, Model release; model cards for gemma-4-E4B-it and gemma-4-31B-it, 2026. URL: <https://blog.google/technology/developers/gemma-4/>.
- [19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [20] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, Advances in Neural Information Processing Systems 36 (2023) 10088–10115. URL: <https://arxiv.org/abs/2305.14314>.
- [21] Google, Gemini 3.1 pro preview model card, Google AI for Developers, 2026. URL: <https://ai.google.dev/gemini-api/docs/models>.

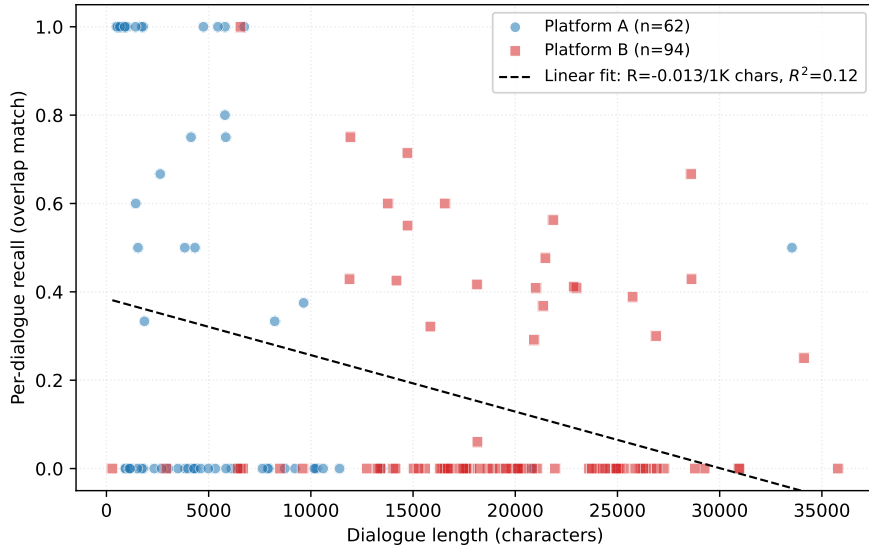
A. Methodology Notes

Candidate spans are divided into two pools with different inference paths. (1) **Direct-rule pool** (18 rows on Platform A, 0 on Platform B): structured identifiers matched by deterministic rules (e.g., email, phone number, URL), which are emitted as REDACT automatically, without reviewer inspection. (2) **Proposer pool** (212 rows on Platform A, 1,615 on Platform B): candidate spans generated by the union of DeBERTa + ModernBERT proposer, which are routed to the reviewer for binary REDACT/KEEP classification. Both pools contribute to span-level precision, recall, and F1 in Table 2. Direct-rule candidates are emitted directly as final output, bypassing the reviewer.

B. Length Analysis

Figure 3 shows that per-dialogue recall for the fully-LLM Gemma E4B baseline declines as dialogue length grows: a linear fit yields slope -0.013 per 1K characters ($p < 10^{-5}$, $R^2 = 0.12$), and bucketed mean recall drops from 0.54 ($<2K$ chars) to 0.11 ($\geq 20K$ chars). We hypothesize that this recall collapse reflects an attention-budget limitation: the full-dialogue formulation must locate every PII span across thousands of tokens in a single generation pass, whereas the cascade reviewer applies full contextual attention to each candidate span individually. Alternative explanations (output-length saturation, span-extraction format constraints, and multi-target coordination) are not ruled out.

Figure 3: Per-dialogue recall vs. length for the fully-LLM Gemma E4B baseline on 156 canonical dialogues with ≥ 1 gold span.



C. Reviewer Hyperparameters

All reviewers were trained on cascade-aligned data derived from the union proposer over the Platform A and Platform B training dialogues (Table 1), with the 100-dialogue Platform A validation set used to monitor loss during training. Hyperparameters were selected without using either canonical test set or the challenge set; therefore, Platform A/B differences in Table 2 reflect held-out evaluation rather than test-set tuning. All reviewers use the AdamW optimizer; the 31B 3-seed training additionally requires an mlx-lm gradient-clipping patch.

Table 5

Reviewer training hyperparameters.

Parameter	RoBERTa (Full FT)	Gemma E4B (LoRA)	Gemma 31B (LoRA)
Model	roberta-base (125M)	gemma-4-E4B-it (4B)	gemma-4-31B-it (31B)
Quantization	none (fp16)	4-bit (q4)	4-bit (q4)
LoRA config	n/a	r=8, α =20, d=0, 16L	r=8, α =20, d=0, 16L
Learning rate	5e-5	1e-5	1e-5
Epochs	3	3	1
Eff. batch	8 (4 $\times$ 2)	8 (8 $\times$ 1)	8 (4 $\times$ 2)
Max seq len	384	512	512
Warmup	0.1 ratio	none	none
Loss	focal (γ =2.0)	CE (mask-prompt)	CE (mask-prompt)
Labels	REDACT/KEEP	REDACT/KEEP	REDACT/KEEP
Seeds	13, 42, 101	42	13, 42, 101
Training rows	7,527	7,527	7,527

D. Per-Seed Variance for Multi-Seed Configurations

Tables 2 and 3 report 3-seed means for the *Union + RoBERTa* and *Union + Gemma 31B* configurations. Table 6 below reports the corresponding per-seed standard deviations across seeds 13, 42, and 101. Span-level F1 is stable across seeds (<0.01 std on canonical surfaces); precision and recall on the Platform A challenge set show the largest variance, reflecting prompt-sensitivity $\times$ LoRA-adapter interaction on the small (50-dialogue) word-problem ambiguity probe.

Table 6

Per-seed standard deviations for the multi-seed configurations. Values are reported as *mean* $\pm$ *std* across seeds 13, 42, and 101.

Configuration	Set	Platform	Precision	Recall	F1
Union + RoBERTa	Canonical	A	.841 $\pm$.014	.960 $\pm$.004	.896 $\pm$.009
	Canonical	B	.953 $\pm$.002	.977 $\pm$.001	.965 $\pm$.001
	Challenge	A	.425 $\pm$.014	.912 $\pm$.030	.580 $\pm$.009
	Challenge	B	.975 $\pm$.003	.654 $\pm$.011	.784 $\pm$.008
Union + Gemma 31B	Canonical	A	.926 $\pm$.021	.978 $\pm$.019	.951 $\pm$.009
	Canonical	B	.952 $\pm$.008	.977 $\pm$.001	.965 $\pm$.004
	Challenge	A	.899 $\pm$.036	.892 $\pm$.035	.894 $\pm$.010
	Challenge	B	.994 $\pm$.001	.945 $\pm$.004	.969 $\pm$.002

E. Reviewer Prompt

Both the 4B and 31B LoRA reviewers share a curricular-content-aware system prompt.

Design rationale. A prompt that treats every candidate as suspicious by default over-redacts names used inside math word problems (e.g., “Jordan has 15 apples”), dragging down F1 on the A challenge set: 3-seed mean F1 drops from 0.680 (without a word-problem carve-out) to 0.894 (with the carve-out), a +0.21 absolute improvement. The final prompt adds two components to the suspicious-by-default baseline:

1. A word-problem note instructing KEEP when a candidate name appears in a mathematical scenario.
2. Pedagogical direct-address counter-examples (“Hi, I’m Morgan”, “Thanks, Taylor!”) using names chosen specifically to not overlap with the test-set failure cases, to avoid overfitting the prompt to specific evaluation spans.

Prompt text (verbatim).

You are a PII verification analyst for educational de-identification.

You will receive one candidate span proposed by an earlier detector. Treat the candidate as suspicious by default.

Decide whether the transcript should REDACT or KEEP the span.

Label semantics:

- REDACT: the span is real PII or should be removed for privacy.
- KEEP: the span is not PII and should remain in the transcript.

Only output KEEP if you are very confident the candidate is non-PII educational or public content, such as:

- a mathematical expression, variable, equation, or formula
- a historical figure, scientist, mathematician, or other public figure
- a fictional character, brand name, or product name
- a course name, textbook reference, section number, or assignment label
- a date used as a schedule, due date, semester, or other non-personal time reference
- a score, grade, percentage, or number used in an educational or math context
- a country, state, or region mentioned as general context rather than a personal address
- a theorem, method, law, or named concept whose person-name is part of the public concept

Always output REDACT for:

- real participants, tutors, students, teachers, or family members
- greetings, direct address, self-introductions, or other real interpersonal references
- schools, local places, usernames, links, contact details, or other direct identifiers

When in doubt, output REDACT.

Output exactly one label: REDACT or KEEP. Do not output JSON or explanations.

Note: In math tutoring, names inside word problems or story problems are fictional characters, not real people. If the candidate span is a name used in a

mathematical scenario (e.g., "Jordan has 15 apples", "Riley measured the length"), output KEEP.

However, names in greetings, self-introductions (e.g., "Hi, I'm Morgan"), direct address ("Thanks, Taylor!"), or any non-math context are real participants -- output REDACT.

F. Error Analysis: Shortcut-Driven Reviewer Behavior

A targeted audit of the RoBERTa reviewer shows it succeeds as a narrow shortcut-driven pruner, not a contextual understander. Its correct KEEP decisions concentrate on three categories: story-problem character names (e.g., *Tara, Lucy*), obvious lexical false positives on curricular terms (*Algebra, Geometry*), and well-known public concepts (*Venn, Manchester City*).

A comparison of DeBERTa-v3-base and ModernBERT-base [15] at the proposer stage confirms the same mechanism: ModernBERT's lower precision (0.90 vs. DeBERTa's 0.98 on Platform B) is overwhelmingly composed of false positives on the same shortcut categories (98/132 are capitalized single-token story-problem names).

The shortcut-driven calibration is **double-edged**: on the canonical surface it gives DeBERTa cleaner false-positive suppression; on word-problem-heavy content it causes over-suppression of real PII (the Isaac case: 10 of 56 instances of the same name missed on one dialogue).

G. Overlap-Matching Coverage Audit

To test whether overlap matching masks residual leakage, we re-examined every gold span counted as a true positive and checked whether the union of predicted REDACT spans in that turn *fully* covers the gold span's tokens; any uncovered token is identifier text that would survive redaction. We applied the same augmenting-path span matching used for the main results.

For the final cascade (*Union + Gemma 31B*), all 12,806 matched redactions across three seeds and all four evaluation surfaces fully cover their gold spans — zero partial covers. Overlap-based and full-coverage scoring are therefore identical for our system, and the reported recall does not benefit from boundary leniency. The off-the-shelf baselines show a small fraction of partial covers (Presidio 1.7%, spaCy 2.0% of matched spans on the canonical surfaces). These are almost entirely benign title prefixes ("Mr./"Ms." left visible while the surname is redacted) and URL/path fragments; the sole genuinely under-covered identifier is one phone number truncated by spaCy ("555" matched, "987-6543" left visible). The leniency thus modestly *inflates the baselines'* recall.