PIIvot: A Lightweight NLP Anonymization Framework for Question-Anchored Tutoring Dialogues

Matthew Zent **Digory Smith** **Simon Woodhead**
Eedi
matthew.zent@eedi.co.uk

Abstract

Personally identifiable information (PII) anonymization is a high-stakes task that poses a barrier to many open-science data sharing initiatives. While PII identification has made large strides in recent years, in practice, error thresholds and the recall/precision trade-off still limit the uptake of these anonymization pipelines. We present PIIvot, a lighter-weight framework for PII anonymization that leverages knowledge of the data context to simplify the PII detection problem. To demonstrate its effectiveness, we also contribute $QATD_{2k}$, the largest open-source real-world tutoring dataset of its kind, to support the demand for quality educational dialogue data.

https://github.com/Eedi/PIIvot
https://huggingface.co/datasets/Eedi/Question-Anchored-Tutoring-Dialogues-2k

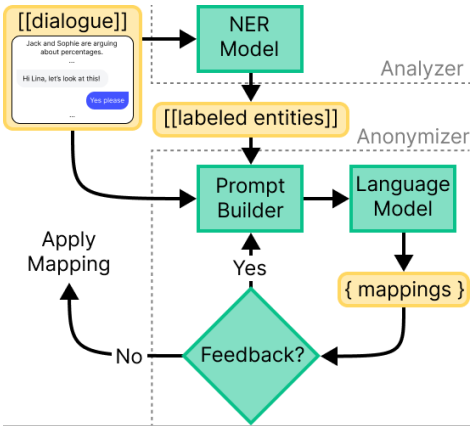

Figure 1: Overview of the PIIvot anonymization framework, which includes a recall-first NER analysis step followed by a context-aware surrogate anonymization step.

1 Introduction and Related Work

Over the past 10 years, we've seen widespread adoption and growth of education technology inside and outside the classroom (Escueta et al., 2017; Manal and Erika, 2024; Manna et al., 2022). Understanding and improving affective learning strategies continues to be one of computing's primary contributions to education research (Mandalapu and Gong, 2019). Among these advancements, high-dosage online tutoring has emerged as a particularly effective intervention to enhance student learning outcomes (Carlana and La Ferrara, 2024; Gortazar et al., 2024), but faces barriers to equitable adoption due to its costs (Aleven et al., 2023).

Large Language Models (LLMs) have been proposed as one way to scale up access (Aleven et al., 2023), but significant challenges persist (Miller and DiCerbo, 2024; Macina et al., 2023b). This rise in evidence-based intelligent systems has fueled demand for high-quality educational data. The few open-source conversational education datasets that exist may not be well-equipped to meet this demand due to their small size (Caines et al., 2020; Wang et al., 2024), degraded quality from crowdworkers (Yu et al., 2017; Stasaski et al., 2020), or reliance on LLM tutors or students (Macina et al., 2023a; Miller and DiCerbo, 2024) which may not be suitable for all downstream tasks (Macina et al., 2023b; Marwala, 2023). Related, mathematical reasoning is a core challenge in generative AI (Rane, 2023), which has seen an influx of reasoning benchmarks to assess and address this limitation (Patel et al., 2021; Li et al., 2021; Gulati et al., 2024). Miller and DiCerbo (2024) and Macina et al. (2025) benchmarks stand out for their focus on these challenges in the context of LLM tutors.

The sensitive nature of student-generated data presents a significant barrier to sharing real-world educational datasets (Hutt et al., 2022). Frequently, research focuses on personally-identifiable information (PII) as the primary challenge of open-science in sensitive contexts (Olatunji et al., 2022). Approaches to data anonymization often grow out of healthcare contexts (Olatunji et al., 2022) and

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27479–27488
November 4-9, 2025 ©2025 Association for Computational Linguistics

generally fall into three categories: limiting access, obfuscation, and minimization. Federated learning limits direct access to records (Antunes et al., 2022; Hutt et al., 2022), but is not suitable for all types of analysis (e.g., qualitative), and is susceptible to de-anonymization attacks (Carlini et al., 2021). Obfuscating PII typically relies on automated recognition (Buchh, 2024; Bosch et al., 2020; Holmes et al., 2023; Singhal et al., 2024) or manual labeling (Miller and DiCerbo, 2024), but identifying PII and overlapping non-PII is challenging even for humans (Singhal et al., 2024). Finally, both data minimization and k-anonymity aim to reduce the risk of data matching by limiting the exposure to and links between identifiable attributes (Ji et al., 2017; Majeed and Lee, 2021; Esfandiari et al., 2022; Sen et al., 2024; Stinar et al., 2024), but may fall short in contexts where entropy is an important metric of dataset quality (Macina et al., 2023a).

Our contribution is two-fold: 1) we developed PIIvot, a novel anonymization framework that reframes PII detection as a simpler potential-PII labeling task and uses an LLM to generate contextually accurate surrogate replacements to preserve data integrity. Using this method, 2) we open-source a large dataset of question-anchored tutoring dialogues ($QATD_{2k}$) from [a large online math learning platform], demonstrating the effectiveness of PIIvot for anonymizing text-based data at scale.

2 Method

2.1 PIIvot

Motivated by the high recall of recent PII identification systems and the persistent challenges they face with precision (Buchh, 2024; Bosch et al., 2020; Holmes et al., 2023; Singhal et al., 2024), we introduce PIIvot, an applied method for text-based anonymization that balances the need to prioritize privacy with data usability. The framework is grounded in two core principles: (1) a recall-first approach to named entity recognition (NER) for identifying potential-PII (Section 2.1.1), and (2) a Hidden-In-Plain-Sight (HIPS) strategy for generating surrogate replacements that preserve text coherence (Section 2.1.2). This process is illustrated in Figure 1. PIIvot is designed as a generalizable framework that can be adapted to different domains and disclosure risks. In the following sections, we detail our specific implementation for transparent data sharing. We acknowledge the inherent tension between reproducibility and privacy preservation in this context. To balance these aims, our full generative pipeline, including all prompts, is publicly available on GitHub, but we reserve the underlying NER model trained on PII-rich data. A working example is available by substituting it for an NER model trained in prior work (Tjong Kim Sang and De Meulder, 2003).

2.1.1 Analysis

The analysis step applies word-level labels to text for named entities that have a risk of containing PII. This methodological choice reflects a decision to minimize the cognitive load of strict PII identification for human annotators (Singhal et al., 2024), but also limits our ability to evaluate performance against similar PII-identification methods (see 5.1). Any suitable NER model can be substituted at this stage, but we caution against openly sharing trained models or open-source details, as they may be used to identify residual PII in the resulting dataset (see 5.1). For $QATD_{2k}$ we used a DeBERTa model fine-tuned on a prior set of 40k labeled student-tutor utterances to label dialogue and question text (See Appendix A.2). [1] Specifically, we label names, locations, URLs, dates of birth, phone numbers, schools, and emails/socials because they are frequent in our data, risk being identifiable, and benefit from granular labels during the anonymization step. Newline characters are replaced with whitespace to avoid out-of-vocabulary tokens. [2] We then apply an IO labeling scheme and first-token aggregation strategy to resolve multi-token predictions into labeled word-level spans. Each message is analyzed in a centered context window that includes both the preceding and following messages in the dialogue. Finally, we automatically clean labeled spans to remove trailing or preceding punctuation to improve the reliability of downstream surrogate replacement.

2.1.2 Anonymization

The anonymization step utilizes labeled spans to generate surrogate replacements under the assumption that the content of non-PII spans can be changed without impacting dataset quality, so long as the same name/location/etc. is consistent

[1] We define an utterance as a single chat message where a talk turn can be made up of one or more consecutive messages.

[2] Initial tests without this preprocessing step resulted in performance degradation on question text, which reflects a generalizability gap for out-of-vocabulary tokens as newlines were absent from training dialogue data since the 'Enter' key corresponds to sending messages on Eedi's chat platform.

throughout the conversation or document. We argue that this assumption holds for $QATD_{2k}$, where the names and locations of word problems are not relevant to the questions' mathematical concepts. This HIPS approach has the added benefit of minimizing the risk of the residual PII problem (Carrell et al., 2013). For labels that can be automatically verified–emails and URLs–we use obfuscation-based anonymization. For $QATD_{2k}$, we use *GPT-4o-2024-11-20* to generate a mapping from the original set of words to an anonymized set, conditioned on the full chat history to ensure replacements are coherent across each dialogue. Each label type includes qualities to preserve in the anonymized text that we include in the prompt (i.e., "When anonymizing [[NAME]], preserve their gender and ethnic background."). Then we apply feedback-based reprompting to enforce measurable qualities of the anonymization (i.e., ensuring the replacement is significantly different from the original).

2.2 Dataset Collection and Processing

Existing conversational tutoring datasets (Macina et al., 2023a; Stasaski et al., 2020; Yu et al., 2017) with annotated talk moves leverage synthetic environments to generate data to scaffold teaching strategies of LLM-based tutors, but limited work has explored these properties in real-world environments. To fill this gap, we curate a dense set of chat-based tutoring sessions on a UK-based learning platform deployed in over 19,000 schools around the world. [3] Conversations are prompted by the student asking for assistance from an on-demand expert tutor while working on a lesson typically assigned by their teacher. We include metadata about the question the student was working on and lesson descriptors. [4]

2.2.1 Initial Filtering

First, we select conversations that started during a Diagnostic Question (DQ), but before an answer was selected. DQs are multiple-choice questions with one correct answer and three incorrect distractors representing common misconceptions. Similarly to Chen et al. (2019), we filter sessions with at least 20 total and 7 messages from either participant, as these sessions are more likely to have meaningful teaching or learning. Then, we filter out US-based students by email domain and school.

Next, we take initial steps to safeguard the tutors and students represented in QATD. We used *omni-moderation-2024-09-26* to filter out conversations with unsafe content. [5] We obtained affirmative consent from 25 of 31 tutors represented in the filtered set because of the high density of individual tutors' conversations. This process resulted in 4,129 dialogues that met our criteria–$QATD_{Candidate}$.

2.2.2 Talk Move Downsampling

Motivated by the growing emphasis on quality over quantity for alignment tasks (Zhou et al., 2023) and data-sharing restrictions, we selectively downsample $QATD_{Candidate}$ to create a dataset that prioritizes diverse examples of tutor talk moves. *Talk moves* are strategies used to support students' mathematical thinking, understanding, and communication (O'Connor et al., 2015). We use the GPT4 talk move classifier from prior work to apply 7 talk move labels (Moreau-Pernet et al., 2024). Because this model was fine-tuned on small group tutoring conversations, we first evaluate its generalizability to 1:1 online tutoring. The first author annotated a weighted stratified sample of 200 tutor messages to conduct a contextual error analysis (see Appendix B) (Chancellor et al., 2023). Except for a systematic error on the 'Getting Students to Relate' label, we see similar performance to the original paper.

To construct our final dataset, $QATD_{2k}$, we first compute TF-IDF scores over talk move labels in $QATD_{Candidate}$, excluding 'None' and 'GSR'. We form $QATD_{2k}$ by greedily selecting dialogues with the max TF-IDF score under two constraints: (1) at most 8 dialogues per distinct DQ, and (2) a maximum of 1000 unique DQs. This strategy results in the most diverse examples of tutoring strategies without oversampling from any single DQ.

2.3 Annotations

We evaluate the performance of PIIvot on $QATD_{2k}$ by manually annotating potential-PII. A codebook was developed during a prior labeling initiative of 40k student-tutor messages and achieved a minimum Weighted F1 score of 0.98 between raters across a subset of 350 dialogues (see Ap-

[3] https://eedi.com/

[4] Questions were originally presented to students as images. The associated text-based metadata was extracted using the Mathpix API v3, then labeled and validated by tutors.

[5] Sexual, sexual/minors, harassment, harassment/threatening, hate, hate/threatening, illicit, illicit/violent, self-harm, self-harm/intent, self-harm/instructions, violence, and violence/graphic.

Dataset	Total Dialogues	Total Turns	Words per Turn		N-Gram Entropy		Turn Uptake	Human			Subject
			Student	Tutor	Student	Tutor		In-Situ	Student	Tutor	
$QATD_{2k}$	1971	**46249**	4.15	14.79	12.74	13.39	0.69	✓	✓	✓	Math
↪ No PIIvot	–	–	4.15	14.80	12.73	13.39	0.69	–	–	–	–
TSCC v2	260	25840	9.91	18.92	13.84	14.48	0.71	✓	✓	✓	Lang.
Bridge	459	2860	2.57	14.98	10.13	11.89	0.74	✓	✓	✓	Math
CoMTA	188	2022	8.32	37.08	11.54	12.07	0.90	✓	✓		Math
CIMA	391	1427	6.58	10.00	8.69	10.36	0.83		✓	✓	Lang.
Burchak	173	2412	3.20	3.47	10.51	10.54	0.59		✓	✓	Lang.
MathDial	2262	29453	37.86	15.88	13.82	13.79	0.84			✓	Math

Table 1: Comparison of available 1:1 tutoring datasets. Uptake is modeled using Demszky et al. (2021). PIIvot had little to no effect on text-based metrics.

Label Set	Source	Precision	Recall	F1
Dialogues	PIIvot	0.984	0.984	0.984
	Annotators	0.993	0.995	0.994
Questions	PIIvot	0.996	0.997	0.996
	Annotators	0.997	0.997	0.997

Table 2: Micro-averaged metrics for potential-PII detection on dialogues and question text compared to ground truth labels.

pendix A.1). The first and second authors and two tutors from the original initiative independently applied the codebook to 68,717 messages and 1000 questions. Discrepancies between the machine and annotator labels were resolved to establish a ground truth. Each dialogue was also flagged for the presence of unsafe content and the absence of a learning event–29 dialogues were removed, 28 for learning and 1 for safety.

3 Results and Discussion

3.1 PIIvot

To assess the effectiveness of the PIIvot framework, we triangulate data from curating $QATD_{2k}$. We report aggregate label metrics to mitigate the small but non-zero risk of residual PII. The high inter-rater reliability observed in the potential-PII labeling task indicates that the task is more straightforward than PII annotation. Table 2 presents macro-averaged metrics comparing our potential-PII NER model against manually annotated labels, evaluated on both dialogue and DQ text. As expected, our model/task outperforms the more challenging PII detection task on student-generated text in comparable educational domains (Buchh, 2024). Table 1 illustrates that PIIvot anonymization has minimal impact on key text characteristics of $QATD_{2k}$. These results present a practical case the PIIvot framework in data-sharing pipelines.

3.2 $QATD_{2k}$

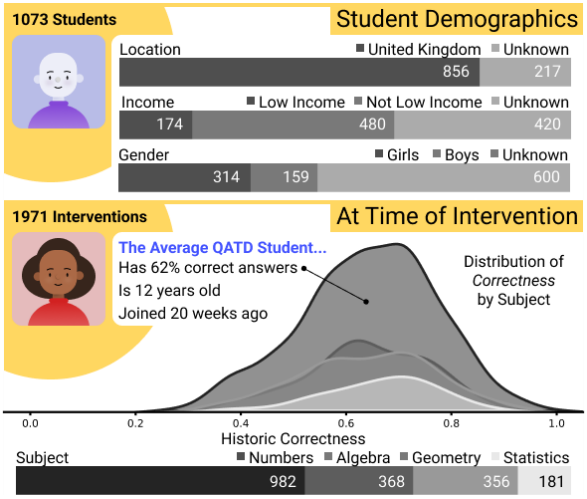

Figure 2: A figure of describing the 1073 students in $QATD_{2k}$. Location, gender, and age are self-reported. The historic correctness plot shows a kernel density estimate (KDE) of student accuracy weighted to prioritize students with 100+ answers.

We shift to a brief reflection on $QATD_{2k}$. Roughly 1% of sessions were removed due to the absence of a learning event, suggesting that talk move downsampling successfully prioritized pedagogically meaningful conversations. Figure 2 presents an overview of the students represented in the data. While experiments on $QATD_{2k}$ are outside the scope of this work, we provide train/test splits to support comparisons across models and methods in future work.

Table 1 situates $QATD_{2k}$ within the broader landscape of available 1:1 tutoring datasets. With more real-world data in the available open-sourced tutoring datasets, two trends emerge. First, LLM tutors/students tend to generate unrealistically long messages. Second, the high uptake metrics of datasets with synthetic tutors–LLMs or crowdsourcing–indicate potential overfitting to stu-

dent turns in a way that diverges from authentic responses. These patterns underscore the importance of real-world tutoring systems to respond effectively in low-information dialogue settings. Future work should include more in situ datasets in benchmark and training data preparation. We open-source $QATD_{2k}$ to support this growing demand for real-world tutoring datasets.

4 Conclusion

We introduce PIIvot, an anonymization framework that balances the trade-off between precision and recall in PII identification, suitable for contexts where the content of overlapping non-PII entities doesn't impact dataset integrity. PIIvot enabled the open sourcing of $QATD_{2k}$ to support future research on effective math tutoring. We present results from curating $QATD_{2k}$ as a practical case for using the PIIvot framework in data-sharing pipelines.

5 Limitations and Ethical Considerations

5.1 Limitations

Our work presents two valuable contributions with the PIIvot framework and the QATD dataset, but both carry important limitations that should be considered in future research and downstream applications.

We acknowledge that the PIIvot framework relies on the assumption that the content of labeled entities is insignificant, which is not true across many domains. Future work should explore improved prompting strategies and/or feedback during anonymization to better preserve the significance of replaced content and mitigate this limitation in new contexts. Recent work demonstrates the potential of incorporating LLM-generated feedback to improve LLM summarization tasks (Song et al., 2025), suggesting a promising direction for PIIvot anonymization feedback. Related, the framework relies on an effective NER that meets the privacy needs of one's data and assumes that these entities are a superset of strict PII. While our decision to only label potential-PII prioritized annotator task load, it limited our ability to assess relative improvements over existing PII detection tools. Additionally, PIIvot uses HIPS to obfuscate PII. We strongly recommend that neither the underlying labels nor the NER models be released alongside datasets, as they may expose residual PII. In our case, there still remains a non-zero risk of residual PII in QATD, despite extensive measures to ensure the safety and privacy of tutors and students. This risk illustrates an inherent limitation of any automated anonymization pipeline and underscores the need to consider a variety of privacy factors outside of identifiably.

Related to the QATD dataset, we highlight four key limitations that reflect trade-offs made to support open-source release and downstream usability. First, we acknowledge that the 'Getting Students to Relate' talk move label may not fully generalize to our 1:1 tutoring context. We include talk move labels in QATD for method transparency, but downstream use of this signal should consider this limitation (Discussed further in Appendix B). Second, this dataset reflects real interactions on Eedi, where tutors occasionally manage multiple students during peak hours and prioritize resolving misconceptions to help students feel confident getting back to their lessons. This context and the reported behavior and demographic factors in Figure 2 should be considered when interpreting tutor and student behavior in QATD. Finally, we acknowledge that our decision to prioritize student privacy by removing student links across tutoring sessions may impact downstream applications of QATD. This decision was made due to the inability to get additional student consent outside of the platform's terms and conditions for the risks conversation linkage could introduce. We underscore that PII anonymization is only one aspect of responsible data sharing and broader privacy concerns.

5.2 Ethical Considerations

This work highlights the range of privacy considerations necessary when open-sourcing data from real educational platforms. While this work is outside the purview of what is traditionally defined as human subjects research, we recognize our responsibility to reflect on its ethical implications–both for dissemination and shaping best practices for future research.

First, Eedi's legal terms of service and privacy policy permit the sharing of personal data with third parties for the purpose of conducting research, but we recognize that legal permission alone is not sufficient. Prior research emphasizes the ethical responsibility of researchers and platform organizers to steward the trust of their users/stakeholders (Commission and others; Zent et al., 2025). Considering these values, we obtained affirmative consent from high-volume contributors, applied data minimization principles to student data, and outline the

following recommendations for appropriate secondary use. In accordance with Eedi's privacy policy, QATD is released for non-commercial research (under cc-by-nc-sa-4.0) aimed at improving student learning outcomes, including, but not limited to, dialogue modeling, model calibration, and tutoring interaction analysis. Attempts to re-identify individuals from QATD are out of scope and violate the intended use of this dataset. We encourage future research to use this dataset to advance understanding of how conversational strategies support learning while upholding these ethical standards.

We further caution researchers to validate third-party APIs used in PIIvot anonymized to ensure prompt inputs are not stored or logged, as they contain unanonymized text. In our case, OpenAI reports not using our prompt data for model training or persistent storage. We encourage future work to consider self-hosting LLMs for highly sensitive contexts.

Finally, we acknowledge the positionally of the authors and annotators of this work as paid employees of Eedi. This relationship carries both privileged access and heightened ethical responsibility. As stewards of users' trust, our proximity to the platform and its data influenced our anonymization decisions. We prioritized safety and privacy, opting for conservative redaction and aggregation strategies and human validation to minimize the risk of re-identification. This commitment reflects our obligation to protect the individuals whose interactions make this research possible.

5.2.1 AI Assistant Disclosure

We used AI assistants, including Copilot and GPT, to support code development and documentation. We used these tools to draft boilerplate code and text for some comments and documentation. All generated content was validated and iterated on to align with our standards.

Acknowledgments

We thank Eedi for supporting this work and the committed tutors whose dedication made this research possible. We are especially grateful to those who contributed their time and expertise to the annotation process.

References

Vincent Aleven, Richard Baraniuk, Emma Brunskill, Scott Crossley, Dora Demszky, Stephen Fancsali, Shivang Gupta, Kenneth Koedinger, Chris Piech, Steve Ritter, Danielle R. Thomas, Simon Woodhead, and Wanli Xing. 2023. Towards the Future of AI-Augmented Human Tutoring in Math Learning. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 26–31, Cham. Springer Nature Switzerland.

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. 2022. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. *ACM Trans. Intell. Syst. Technol.*, 13(4). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Nigel Bosch, R. Wes Crues, Najmuddin Shaik, and Luc Paquette. 2020. "Hello, [REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums. In *Educational Data Mining*.

Irshad A Buchh. 2024. Enhancing PII Detection in Student Essays: A Longformer-based Approach with Synthetic Data Augmentation. In *2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pages 143–149.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The Teacher-Student Chatroom Corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.

Michela Carlana and Eliana La Ferrara. 2024. Apart but connected: Online tutoring, cognitive outcomes, and soft skills. Technical report, National Bureau of Economic Research.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Stevie Chancellor, Jessica L. Feuston, and Jayhyun Chang. 2023. Contextual Gaps in Machine Learning for Mental Illness Prediction: The Case of Diagnostic Disclosures. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Guanliang Chen, Rafael Ferreira, David Lang, and Dragan Gasevic. 2019. Predictors of Student Satisfaction: A Large-Scale Study of Human-Human Online Tutorial Dialogues. *International Educational Data Mining Society*. Publisher: ERIC.

Federal Trade Commission and others. Protecting Consumer Privacy in an Era of Rapid Change—Recommendations for Businesses and Policymakers, FTC Report, Mar. 2012.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 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 (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Maya Escueta, Vincent Quan, Andre Joshua Nickow, and Philip Oreopoulos. 2017. Education technology: An evidence-based review. Publisher: National Bureau of Economic Research.

Hossein Esfandiari, Vahab Mirrokni, and Jon Schneider. 2022. Anonymous bandits for multi-user systems. *Advances in Neural Information Processing Systems*, 35:12422–12434.

Lucas Gortazar, Claudia Hupkau, and Antonio Roldán-Monés. 2024. Online tutoring works: Experimental evidence from a program with vulnerable children. *Journal of Public Economics*, 232:105082.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-AXIOM: A Functional and Static Benchmark for Measuring Higher Level Mathematical Reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Langdon Holmes, Wesley Morris, Harshvardhan Sikka, and Anne Trumbore. 2023. Deidentifying Student Writing with Rules and Transformers. pages 708–713.

Stephen Hutt, Ryan S. Baker, Michael Mogessie Ashenafi, Juan Miguel Andres-Bray, and Christopher Brooks. 2022. Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data. *British Journal of Educational Technology*, 53(4):756–775. Publisher: John Wiley & Sons, Ltd.

Shouling Ji, Prateek Mittal, and Raheem Beyah. 2017. Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey. *IEEE Communications Surveys & Tutorials*, 19(2):1305–1326.

Wenda Li, Lei Yu, Yuhuai Wu, and Lawrence C. Paulson. 2021. IsarStep: a Benchmark for High-level Mathematical Reasoning. In *International Conference on Learning Representations*.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. _eprint: 2502.18940.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and Challenges in Neural Dialog Tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Abdul Majeed and Sungchang Lee. 2021. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, 9:8512–8545.

Hamarsha Manal and Kopp Erika. 2024. EduTech Revolution: The Dynamic Role of ICT in Shaping Learning Environments.

Varun Mandalapu and Jiaqi Gong. 2019. Understanding Affective Dynamics of Learning Toward a Ubiquitous Learning System. *GetMobile: Mobile Comp. and Comm.*, 23(2):9–15. Place: New York, NY, USA Publisher: Association for Computing Machinery.

Manpreet Singh Manna, Balamurugan Balusamy, Kiran Sood, Naveen Chilamkurti, and Ignisha Rajathi George. 2022. Edutech Enabled Teaching: Challenges and Opportunities. Publisher: CRC Press.

T Marwala. 2023. Algorithm bias—synthetic data should be option of last resort when training ai systems. *United Nations University*.

Pepper Miller and Kristen DiCerbo. 2024. LLM Based Math Tutoring: Challenges and Dataset.

Baptiste Moreau-Pernet, Yu Tian, Sandra Sawaya, Peter Foltz, Jie Cao, Brent Milne, and Thomas Christie. 2024. Classifying Tutor Discursive Moves at Scale in

Mathematics Classrooms with Large Language Models. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, pages 361–365, New York, NY, USA. Association for Computing Machinery. Event-place: Atlanta, GA, USA.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Catherine O'Connor, Sarah Michaels, and Suzanne Chapin. 2015. "Scaling Down" to Explore the Role of Talk in Learning: From District Intervention to Controlled Classroom Study. In *Socializing Intelligence through Talk and Dialogue*, pages 111–126.

Iyiola Olatunji, Jens Rauch, Matthias Katzensteiner, and Megha Khosla. 2022. A Review of Anonymization for Healthcare Data. *Big Data*, 12.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Nitin Rane. 2023. Enhancing mathematical capabilities through ChatGPT and similar generative artificial intelligence: Roles and challenges in solving mathematical problems. *Available at SSRN 4603237*.

Aniruddha Sen, Christine Task, Dhruv Kapur, Gary Howarth, and Karan Bhagat. 2024. Diverse community data for benchmarking data privacy algorithms. *Advances in Neural Information Processing Systems*, 36.

Shreya Singhal, Andres Felipe Zambrano, Maciej Pankiewicz, Xiner Liu, Chelsea Porter, and Ryan S. Baker. 2024. De-Identifying Student Personally Identifying Information with GPT-4. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 559–565, Atlanta, Georgia, USA. International Educational Data Mining Society.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. Learning to Summarize from LLM-generated Feedback. _eprint: 2410.13116.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A Large Open Access Dialogue Dataset for Tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Frank Stinar, Zihan Xiong, and Nigel Bosch. 2024. An Approach to Improve k-Anonymization Practices in Educational Data Mining. *Journal of Educational Data Mining*, 16(1):61–83. Section: EDM 2024 Journal Track.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. 2017. The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Matthew Zent, Seraphina Yong, Dhruv Bala, Stevie Chancellor, Joseph A Konstan, Loren Terveen, and Svetlana Yarosh. 2025. Beyond the Individual: A Community-Engaged Framework for Ethical Online Community Research. *arXiv preprint arXiv:2503.13752*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Potential-PII NER Model

In this section, we describe our process for NLP model fine-tuning on the potential-PII classification task. We developed our own classification model for two reasons: 1) initial exploration of existing PII identification models revealed poor performance on UK names, and 2) we wanted more control of label granularity to support surrogate replacement. First, we outline the annotation process to support supervised fine-tuning, and then we discuss our experimental setup and hyperparameters.

A.1 Potential-PII Annotation

Annotation for an independent batch of 66,059 tutor/student messages took place from May to August 2024 by paid employees of Eedi. The first and second authors (US/UK/Male) and 4 expert tutors (UK/Female) participated in this annotation process; tutors annotated messages while not actively helping students. We used the open-source

DeBERTa-PIIvot-NER-IO	
Precision	0.93
Recall	**0.98**
F1	0.94
Balanced Accuracy	0.98

Table 3: Performance of the final DeBERTa-PIIvot-NER-IO model on a held-out test set. Macro scores are computed over positive labels; balanced accuracy includes the 'O' (non-PII) class.

annotation tool Doccano to apply labeled spans to tutor messages (Nakayama et al., 2018). Before manual annotation, labels are prepopulated using a regex applied matching strategy using the known first and last names of the tutor and student, as well as common word problem names.

First, we developed and validated a codebook to support potential-PII labeling. Annotators independently labeled a subset of 350 dialogues, achieving a minimum Weighted F1 score of 0.98 between raters. This high level of agreement indicated that the codebook was well calibrated, and no significant changes were needed. The codebook, including annotation instructions, is available in the PIIvot repository. Annotators individually applied the codebook to the remaining messages to support model fine-tuning.

A.2 Model Fine-Tuning

We conduct our model fine-tuning experimentation on a single NVIDIA Tesla V100 GPU using deberta-v3-base (184M parameters) and bert-base-uncased (110M parameters) (He et al., 2021; Devlin et al., 2018). To support model testing, we use stratified sampling on the minority label for a given message to generate train (64%), test (20%), and validation (16%) splits (see Table 4 for approximate label splits). We initialize a sequential set of hyperparameter grid searches over a select subset of approaches. This greedy approach allowed us to explore a wide variety of modeling approaches without ballooning compute time.

Optimal configurations are shown in bold. For each search, we also include two learning rates (1e$-$5 vs. **2e$-$5**) and two labeling schemas (**IO** vs. IOB2) over 4 epochs with early stopping on performance degradation. In order we test BERT vs. **DeBERTa**, Adam vs. **AdamW**, **raw + synthetic** vs. raw data, and **windowing** vs. non-windowing. For the synthetic data, 20 PII-rich synthetic student-tutor conversations were manually created by the authors to augment the training and validation data with examples of imbalanced classes. For the windowing condition, model inputs include the prior and preceding message. In total, fine-tuning and final model training required 36 GPU hours.

The final model was trained using the AdamW optimizer with β_1=0.9, β_2=0.999, ϵ=1e$-$8, and a weight decay of 0.01. We used a learning rate of 2e$-$5, a batch size of 4, and trained for 3 epochs with a random seed of 42. We report model performance on the hold-out test split in Table 3.

B Talk Move Classification

To facilitate downsampling, talk moves were applied using the GPT-based classification model in Moreau-Pernet et al. (2024). Labels include: 'Pressing for accuracy', 'Keeping everyone together','Revoicing', 'Restating', 'Pressing for reasoning', 'Getting students to relate to another's ideas', and 'None'. The model was fine-tuned on conversation transcripts from small-group math tutoring sessions. Both sets of authors decided use of this artifact was acceptable so long as performance was validated in this new context.

B.1 Contextual Error Analysis

To evaluate the generalizability of the model for 1:1 chat-based math tutoring sessions, the first author manually annotated a validation set of 200 tutor utterances sampled through weighted stratified sampling using the original codebook of Moreau-Pernet et al. (2024). The sample distribution was flattened by 0.8 of the original label distribution represented in $\text{QATD}_{\text{Candidate}}$ in order to validate more examples of minority class labels (see Table 5).

The first author conducted a contextual error analysis on all mismatched labels (Chancellor et al., 2023). This method introduces qualitative coding and thematic analysis into traditional ML error analysis to understand contextual details missed in annotation tasks. We adopt contextual error analysis for this task because it is well-equipped to reveal aspects of the model that don't generalize to 1:1 tutoring contexts.

We begin by qualitatively coding tutor messages and memoing contextual errors. Two themes emerged from these artifacts that we use to describe the errors introduced by applying the model to this new context. First, a small source of errors

Label	Description	Approximate Support		
		Train	Validation	Test
I-date_of_birth	Birth date detail	90 (94%)	20 (95%)	<10 (0%)
I-email_social	Email address, social media handle, or profile	80 (92%)	20 (95%)	<10 (0%)
I-location_address	Geographical detail indicative of a person's location	100 (65%)	30 (69%)	10 (0%)
I-name	A person's full, partial, or nickname	2300 (0%)	600 (0%)	700 (0%)
I-phone_number	Phone number	80 (96%)	20 (95%)	<10 (0%)
I-school_name	School name	70 (95%)	20 (94%)	<10 (0%)
I-url	URL	100 (58%)	30 (68%)	20 (0%)

Table 4: IO label schema and approximate support (with % synthetic) in each dataset split.

Label	Support $QATD_{Candidate}$	Support Validation Set	Support Original	F1	F1 Excl. <GSR>	F1 (Original)
<None>	0.42	0.28	0.73	0.8991	0.8991	**0.96**
<Keep Together>	0.36	0.20	0.09	0.8333	**0.8750**	0.81
<Revoicing>	0.12	0.18	0.03	0.8986	**0.8986**	0.76
<Press for Accuracy>	0.06	0.16	0.13	0.7733	0.8286	**0.88**
<Getting Students to Relate>	0.02	0.08	0.004	0.0000	–	**0.75**
<Press for Reasoning>	0.002	0.06	0.006	0.7857	**0.9565**	0.94
<Restating>	0.0003	0.04	0.008	0.8000	0.8000	**0.95**

Table 5: Distribution and F1 scores for talk move labels comparing our dataset with the original metrics in Moreau-Pernet et al. (2024) F1 scores reported with and without the <GSR> label.

related to the **multi-message chat turns** present in QATD. When tutors span their intent across multiple messages, the temporal fragmentation leads to label mismatches or partial crediting of complex moves. A major class of errors stemmed from the **fictional argument questions** used in DQs. These items frame math problems as debates between two fictional students, and tutors frequently probe the student to reason about the validity of each claim. While these prompts closely resemble <Getting Students to Relate> (<GSR>) in structure, the original codebook doesn't take a stance on whether this label applies to a fictional setting. We chose not to apply the <GSR> to these instances, but acknowledge this is a gray area for a clearly out of context example. We note that in all cases, <GSR> was associated with another positive talk move label. We use these observations to motivate reporting a second set of metrics excluding the controversial label. We find that our results are comparable to those reported in Moreau-Pernet et al. (2024) and use this as grounds for applying the classifier to downsample $QATD_{Candidate}$ to $QATD_{2k}$.