EduCoder: An Open-Source Annotation System for Education Transcript Data

Guanzhong Pan[1], Mei Tan[2], Hyunji Nam[2], Lucía Langlois[2], James Malamut[2], Liliana Deonizio[2], Dorottya Demszky[2]

[1]Carnegie Mellon University, [2]Stanford University

arthurp@andrew.cmu.edu, {mxtan, hjnam, lucilang, jmalamut, lilianas, ddemszky}@stanford.edu

arXiv:2507.05385v3 [cs.CL] 11 Aug 2025

Abstract

We introduce EduCoder, a domain-specialized tool designed to support utterance-level annotation of educational dialogue. While general-purpose text annotation tools for NLP and qualitative research are abound, few address the complexities of coding education dialogue transcripts—with diverse teacher-student and peer interactions. Common challenges include defining codebooks for complex pedagogical features, supporting both open-ended and categorical coding, and contextualizing utterances with external features, such as the lesson's purpose and the pedagogical value of the instruction. EduCoder is designed to address these challenges by providing a platform for researchers and domain experts to collaboratively define complex codebooks based on observed data and incorporates both categorical and open-ended annotation types along with contextual materials. Additionally, it offers a side-by-side comparison of multiple annotators' responses, allowing comparison and calibration of annotations with others to improve data reliability. The system is open-source[1] with demo video available[2].

1 Introduction

Teacher-student and student-student dialogue play a critical role in learning, offering insight into instruction quality, student reasoning and understanding (Mercer and Howe, 2012; Michaels et al., 2008). In recent years, the release of an increasing number of conversational datasets reflects growing interest in analyzing such dialogues (Stasaski et al., 2020; Suresh et al., 2022; Demszky and Hill, 2023; Wang et al., 2024; Macina et al., 2023). Since manual analysis is labor-intensive and difficult to scale, there has been growing efforts to develop natural language processing (NLP) based measures of educationally relevant constructs in transcript data (D'Mello and Graesser, 2012; Jensen et al., 2021; Hunkins et al., 2022; Demszky et al., 2021; Alic et al., 2022; Suresh et al., 2018; Zhao et al., 2024). However, progress in NLP model development for analyzing teachers' and students' talk moves hinges on rigorous and thoughtful human annotation, critical for model validation and training.

Challenges & Motivation. Rigorous human annotation of educational dialogue is nontrivial, reconciling literature-based learning theories, practitioner expertise, linguistic features, and contextual information. It follows an iterative calibration process where researchers and practitioners collaboratively specify annotation codebook. In addition to conceptual factors, there are multiple practical factors making process quite challenging.

First, **no purpose-built, adaptive tool** exists for annotating educational transcripts with iterative co-design and calibration support. Most existing platforms target general NLP taggings like named entity recognition (NER) or part-of-speech (POS) tagging, and lack support for annotating high-inference pedagogical features such as the discourse moves of teachers and students–and the latent cognitive processes they enact (Nakayama et al., 2018; Honnibal and Montani, 2017; Perry, 2021). Qualitative analysis platforms (ATLAS.ti Scientific Software Development GmbH, 2023; Dhakal, 2022) support open coding but lack support for structured, utterance-level annotation and real-time analysis for collaborative calibration. As a result, many researchers rely on spreadsheets or ad hoc solutions, which are time-consuming to build, often clunky to use, and are poorly integrated with modern machine learning workflows.

Second, while non-technical annotators—such as educators and qualitative researchers—may possess deep domain expertise, they often face **signifi-**

[1]https://github.com/EduNLP/EduCoder

[2]https://www.youtube.com/watch?v=5YFR2seN0_k&t=4s

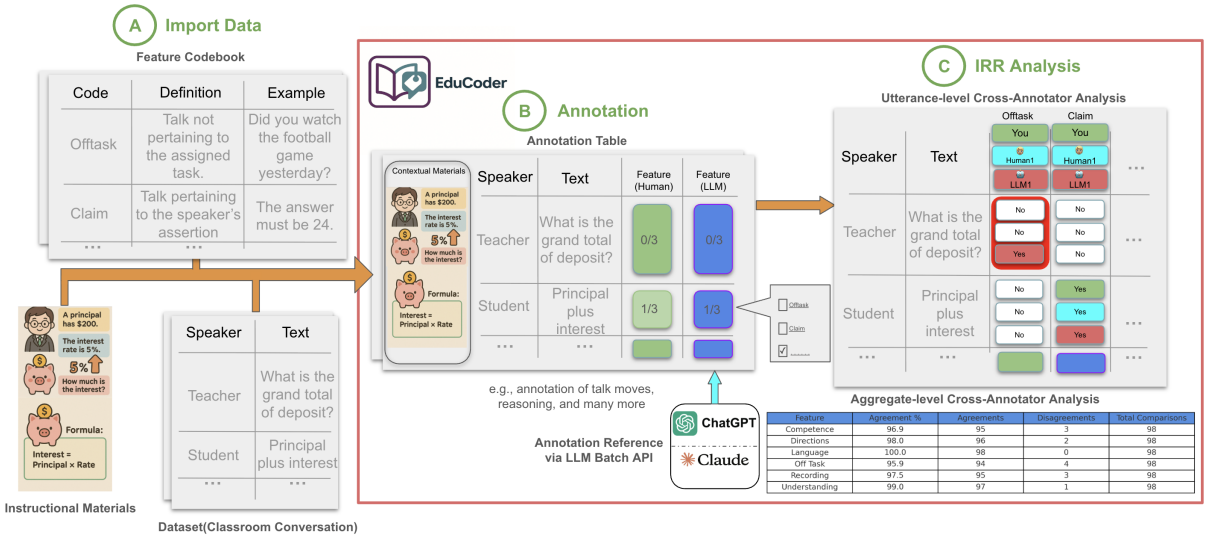


Figure 1: Overview of EduCoder. EduCoder is designed to facilitate the collaborative annotation of educational dialogue data. It is a comprehensive, end-to-end platform for A. importing and preprocessing conversation transcripts, B. collaborative utterance-level annotation with customizable codebooks, and C. real-time inter-rater reliability (IRR) monitoring and cross-annotator comparison. This system aims to enhance the usability and collaboration of educational dialogue research.

cant barriers when using computational annotation tools. These challenges extend beyond a steep learning curve to include usability issues, interface complexity, and inefficiencies in accessing dispersed contextual information. Although conventional tools like spreadsheets or manual methods may be more accessible, they often lack the integrated features necessary for high-quality analysis—such as support for cross-annotator comparison, critical for effective annotator training and reliable results. Additionally, they typically do not provide a user-friendly interface that consolidates all relevant contextual materials into a single, centralized workspace for improving annotation efficiency and supporting annotator understanding.

Third, there is growing interest in **leveraging large language models (LLMs) as annotation tools** to help meet the demand for high-quality labeled data. While LLMs have shown promising performance on tasks such as sentiment analysis, political stance detection, and educational content classification—sometimes achieving IRR comparable to that of humans in well-defined domains (Bojić et al., 2025; Long et al., 2024)—their reliability remains less understood when applied to more complex or socially nuanced constructs. Recent work has proposed treating LLMs as additional "raters" to help quantify and control measurement error in human annotation processes (Broska et al., 2025; Calderon et al., 2025). Building on this line of research, we take it a step further: rather than treating LLMs as standalone annotators, we incorporate their outputs as one of several points of reference during human annotator calibration. This approach allows us to explore how LLMs might support, rather than replace, human judgment in complex annotation tasks.

Our System EduCoder addresses these challenges through an intuitive interface designed for utterance-level annotation in educational dialogue analysis. The system supports:

- **Customizable Codebook:** Researchers can upload their own codebook of features, allowing for tailored annotation that aligns with different research questions.
- **In-text Annotation & Contextual Materials:** The tool features an intuitive, side-by-side interface that enables annotators to apply codes directly to individual transcript utterances. To facilitate contextually grounded and accurate annotation, the interface presents the full transcript alongside relevant contextual materials, including task instructions and screenshots of the questions students were engaging with. Codebook references are integrated directly into the interface to provide clear definitions of features and avoid confusion along with illustrative examples and non-examples.
- **Calibration Support:** EduCoder offers built-

in diagnostics for IRR (Cohen, 1960), highlights disagreements to facilitate calibration, and enables comparison with other annotators to support iterative refinement of codebooks. In addition, the tool integrates annotations generated by LLMs, such as ChatGPT-4o (OpenAI, 2024) and Claude 4 Sonnet (Anthropic, 2025), to facilitate human-AI collaboration in annotation. LLM annotations can serve as an immediate reference point for human coders and support calibration. By comparing their judgments with those of other humans and LLMs, annotators can identify discrepancies, refine interpretations, and enhance their understanding of the coding task.

To demonstrate EduCoder's flexible design and ensure its accessibility, we provide an online playground[3] with examples, including sample transcripts and a codebook, drawn from real-world classroom dialogue. We invite the community to use EduCoder, adapt our provided materials, and contribute to a growing ecosystem of tools for advancing educational dialogue research.

2 Related Works

The NLP community has benefited greatly from general-purpose annotation toolkits like INCEpTION (Klie et al., 2018), Prodigy (Montani and Honnibal), Label Studio (Tkachenko et al., 2020-2022), and Doccano (Nakayama et al., 2018), which standardize data labeling tasks such as NER and text classification. The newer versions of Label Studio and Prodigy also feature LLM integrations for machine coding. However, they are limited in their ability to support collaborative, iterative codebook refinement and the nuanced yet structured open-ended coding of transcripts necessary for education dialogue research. For instance, transcript annotation is often conducted at the utterance level while preserving their sequential nature—this task is ill-suited to the format of most of these tools.

Recent tools like AWOCATo (Daudert, 2020), LightTag (Perry, 2021), POTATO (Pei et al., 2022), EASE (Deng et al., 2023), AutoDive+ (Wang et al., 2025), offer more advanced features for annotation tasks to facilitate research. AWOCATo offers customizability through JSON configurations, including free-text and categorical labeling. LightTag features annotator productivity enhancements (e.g. batching, pre-annotation) and inter-annotator agreement metrics. POTATO offers extensive customization (through various task templates), active learning, and quality control mechanisms. EASE integrates multi-task active learning (utilizing annotator information) and direct integration with LLMs. AutoDive+ supports multimodal annotation, active learning, and a community-driven model repository. Despite these strengths, these tools do not adequately facilitate iterative, collaborative codebook refinement and real-time IRR metrics; nor do they offer extensive contextual referencing needed to preserve the sequential nature of transcripts and account for additional contextual information.

Qualitative Data Analysis (QDA) tools like NVivo and ATLAS.ti support interpretive coding and are well suited for qualitative education dialogue analysis. However, these tools lack support for structured utterance-level annotations essential for computational modeling and real-time IRR feedback that can facilitate annotator calibration.

Further, the integration of human judgment with AI systems has gained significant attention in NLP, particularly in reframing human-AI collaboration for explanatory tasks (Wiegreffe et al., 2021). Recent work has explored how LLMs parse and interpret text compared to human annotators (Hong et al., 2023), highlighting the importance of understanding differences in human versus AI text processing patterns (Muñoz-Ortiz et al., 2023). This research informs our approach to incorporating LLM-generated annotations as a reference while maintaining human expertise as the primary decision-making authority.

EduCoder sits at the intersection of NLP, qualitative research, and learning analytics, aiming to unify a fragmented software ecosystem. By integrating customizable codebook, real-time IRR feedback, and human-AI comparison into an accessible system, EduCoder is designed to bridge these methodological divides, supporting the multi-faceted nature of educational interaction research.

3 Design Principles

EduCoder follows these core principles:

I. Minimalistic Data Structure. EduCoder requires only that the transcript include identifiable *speaker* and *text* columns, and that the codebook contain distinct *code* and *definition* columns. It also supports additional columns for usability (e.g., *example* columns in the codebook and *segment* columns indicating sections of the transcript).

II. Iterative Refinement and Data Reuse. To sup-

[3] https://educoder.vercel.app/

Feature	EduCoder	INCEpTION	Prodigy	Doccano	Label Studio	LightTag	POTATO	AWOCATo	EASE	AutoDive+
Custom Codebook import	✓								✓	
Utterance-level annotation	✓	–	–	–	–	–	–	✓	✓	✓
Real-time IRR & visual diff	✓	–				–	–	–	–	
Instructions & Contextual Materials	✓						✓		✓	✓
LLM for Annotation	✓		–	–	–	–	–	–	–	✓
Usability for non-technical users	✓	–	–	✓	✓	✓	✓	✓	✓	✓
Open-source / Free for research	✓	✓	×	✓	✓	✓	Free for Academia	✓	✓	–
License	MIT	Apache 2.0	Commercial	MIT	Apache 2.0	MIT	Polyform Shield	GPL	GPL	Request only

Table 1: Comparison of EduCoder with widely used annotation tools across key capabilities relevant to educational dialogue annotation. A checkmark (✓) indicates the tool fully supports the feature. A dash (–) indicates the tool provides the functionality at a basic or partial level but does not meet the specific requirements. A blank cell indicates the feature is not supported.

port the iterative nature of qualitative analysis (Weston et al., 2001; Dunivin, 2025) and the complexity of specifying nuanced social and pedagogical constructs, EduCoder allows for import and export of annotated transcripts. Researchers can load previously annotated transcripts back into the system to track how coding patterns evolve across different annotators or over time. This design enables analysis of annotation trends, distinguish between task ambiguity rooted in the codebook from systematic differences in the perspectives of individual annotators. The system incorporates established principles from case study research methodology (Eisenhardt, 1989) to support theory-building through systematic data collection and analysis.

III. Accessibility for Non-Technical Users. Usage of such computational tools should not require programming expertise to contribute domain knowledge (Park et al., 2021; Garlan et al., 2012). EduCoder provides an intuitive, checkbox-style interface with contextual reference materials and rubric definitions displayed alongside transcript views. The interface is designed to reduces cognitive load, enhance tasking understanding, and enable cross-discipline collaboration with domain experts such as educators and education researchers.

4 EduCoder

4.1 System Architecture

EduCoder supports two primary user roles: (1) *Research Administrators* who manage codebook, project settings, and export results, and (2) *Annotators* who perform the actual coding work. Each role has tailored interfaces that balance functionality with usability.

4.2 Annotation Workflow

Codebook Upload and Configuration. Researcher administrators can import custom codebooks that define constructs to be annotated. They can include high-level feature categories, codes, and detailed definitions with examples (and non-examples) that appear contextually during annotation to provide clear guidance to annotators in real-time and minimize labeling errors.

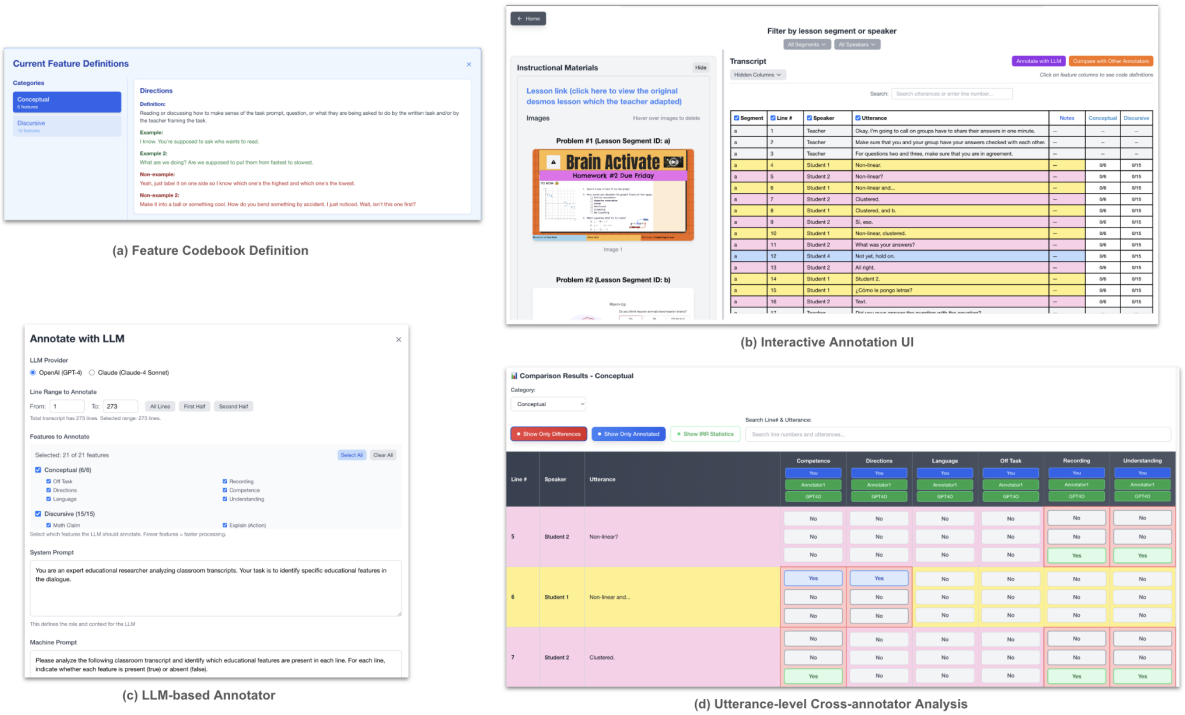
Transcript Import and Preprocessing. Educational transcripts can be uploaded in XLSX or CSV format with required speaker and utterance text columns, plus optional metadata such as timestamps. The system can also segment transcripts (e.g., to indicate a change in learning activity) based on a given *segment* column, filter by keywords/speakers and presents them in a clean interface that preserves conversational flow while enabling utterance-level annotation.

Interactive Annotation Interface. The core experience centers on a dual-pane layout with transcripts alongside feature selection panels. Annotators use checkbox interactions to mark the presence of each code, with definitions available on click.

Real-time IRR Monitoring. As annotations are completed, EduCoder computes IRR metrics both overall and per feature category, with visual indicators showing low agreement features to help identify systematic disagreements or codebook ambiguities. Such feedback accelerates the iterative codebook refinement and annotator calibration processes described by qualitative coding methodologies (Weston et al., 2001), enabling smaller batches and frequent comparison. The system implements established IRR measures, including Cohen's kappa (Cohen, 1960) and Krippendorff's alpha (Krippendorff, 2018), used for ensuring consistency in qualitative research (Hallgren, 2012; Hayes and Krippendorff, 2007).

4.3 Deployment and Accessibility

EduCoder can be deployed both locally or exposed publicly online. Each running instance of EduCoder serves one task with unified features. Data owners (e.g., schools or research teams) can execute private deployments for sensitive data via GitHub, while the public instance supports collab-



(a) Feature Codebook Definition

(b) Interactive Annotation UI

(c) LLM-based Annotator

(d) Utterance-level Cross-annotator Analysis

Figure 2: Screenshots of example tasks supported by EduCoder. Example 2a shows the codebook definition interface for browsing feature descriptions and examples. Example 2b shows the interactive annotation interface for labeling transcript utterances. Example 2c shows the LLM configuration panel for selecting features, line ranges, and prompt templates. Example 2d shows the utterance-level cross-annotator comparison interface for visualizing agreement across multiple annotators or models.

orative projects across different annotators.

The platform is released under an MIT license, ensuring broad accessibility while encouraging community contributions. We also include extensive documentation such as video tutorials, an example codebook, and step-by-step guides for common workflows. The goal is to support lowering technical barriers for educational researchers.

5 Evaluation and Case Study

To evaluate EduCoder's usability and effectiveness, we conducted a focus group at a university summit for educators interested in the application of language technology in education. The study involved 16 K-12 mathematics teachers and two university-based math education professors. Participants completed a 45-minute training session followed by a 60-minute collaborative annotation task using common classroom dialogue transcripts. The codebook included both categorical features (e.g., "is the student playing the role of xyz?") and free-response fields capturing the annotator's observations of students' mathematical understanding and their relevance to the lesson's purpose. Our evaluation methodology follows established principles for credibility in qualitative analysis (Patton, 2002) and best practices for coding in-depth interviews (Campbell et al., 2013).

The annotation process involved calibration exercises where participants annotated one to three reference transcripts to establish baseline agreement before independent coding. This small set of transcripts served as reference points for calibration during the independent annotation sessions. Annotators compared their interpretations of the codebook on the new transcripts against these reference examples to maintain consistent coding practices.

After the annotations were completed, all participants provided positive feedback about the interface and especially emphasized the ease of annotating diverse types and the value of real-time comparison with other annotators. This also prompted meaningful questions during the calibration discussions, as annotators could easily reference others' responses and discuss whether or not they agreed with them. Teachers found the side-by-side comparison view particularly helpful for resolving disagreements and aligning on the codebook. Participants also appreciated the notes for open-ended feedback, as they could share nuanced observations

that the categorical features could not fully capture.

A key challenge was calibrating different annotators' interpretations of the open-ended codes. This is a well-known issue, often described as "complex and iterative" in qualitative research (Cole, 2024). While annotators found it easier to agree on binary labels, some of the more nuanced codes that required contextualization led to longer discussions to determine whether an observed instance represented a positive or negative example of the label. EduCoder's visual feature highlighting the disagreement between the annotated labels were essential in identifying areas of confusion and thereby leading to targeted calibration discussions. Facilitators leveraged this feature during calibration sessions to help guide conversations. These tools encouraged participants to reflect on their own interpretations of the codes by comparing with others, and contribute to the collaborative building of the coding rubric. Several teachers suggested adding a "flag for discussion" feature to bookmark ambiguous cases for team review. The professors emphasized that surfacing subtle interpretive differences is critical for training reliable annotators and developing robust qualitative coding schemes.

Potential use-cases in teacher training and instruction reflections. The study also highlighted the potential for EduCoder to support practitioner reflection and co-design. When reviewing their own classroom conversations, teachers reported that comparing their own analysis with the LLM-generated annotations provided a valuable external perspective. This process sparked discussions about what different features of student talk the LLM was sensitive to, versus what they, as practitioners, noticed. This feedback loop suggests the tool's utility not only for transcript annotation but also for professional development and the co-creation of research frameworks that bridge pedagogical expertise and computational analysis. The integration of LLM references address recent developments in evaluating creative and linguistic patterns in human versus AI-generated content (Ismayilzada et al., 2024), providing users with multiple perspectives on their coding decisions.

Challenges with machine annotations. However, our findings revealed that LLM annotations displayed notably low agreement with human across multiple pedagogical features. This observed challenge aligns with the prior claim (Patton, 2002) that coding is a complex process built on active annotator interpretation and contextual understanding. Our case study suggests that while LLMs have shown capabilities across different domains, they should not be fully trusted or used to replace expert human labelers. LLM-generated annotations may complement human annotations and serve as useful reference points for prompting calibration among annotators; however, our case study shows that they cannot fully replace the interpretative work of human annotators – especially in educational contexts, where meaningful annotations require deep pedagogical content knowledge and understanding of student's learning trajectories.

6 Conclusion

We introduce EduCoder, an open-source annotation platform designed to enhance the study of educational dialogue. Freely available under an MIT license, it offers a user-friendly system complete with documentation, tutorials, and example materials. EduCoder incorporates best practices for qualitative analysis by supporting customizable codebooks and utterance-level coding with context. It additionally supports cross-rater comparison, incorporating LLM judgements. By providing an accessible, collaborative environment, EduCoder unifies research efforts, bridging the gap between pedagogical expertise and computational analysis to improve teaching and learning.

7 Limitations and Future Work

While EduCoder addresses several critical gaps in educational annotation tools, the current system has important limitations that constrain its applicability and suggest directions for future development.

Open-Ended Note Analysis. While the system collects free-form notes, systematic comparison across annotators remains manual and labor-intensive. Future work will implement automated analyses of open-ended notes (e.g., via topic modeling) to identify patterns in annotator responses.

Feature Category Extension. Another useful extension would be to support additional annotation types beyond categorical and free-response formats, such as Likert scales and preference labels, to better accommodate the diverse needs of annotators.

References

Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 224–233.

Anthropic. 2025. Claude 4 sonnet model card. System card published May 2025; training data cut-off: March 2025; accessed July 2 2025.

ATLAS.ti Scientific Software Development GmbH. 2023. ATLAS.ti Mac (version 23.2.1) [qualitative data analysis software]. Accessed: 2025-07-04.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaitė, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Evaluating large language models against human annotators in latent content analysis: Sentiment, political leaning, emotional intensity, and sarcasm. *arXiv preprint arXiv:2501.02532*.

David Broska, Michael Howes, and Austin van Loon. 2025. The mixed subjects design: Treating large language models as potentially informative observations. *Sociological Methods & Research*, page 00491241251326865.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970*.

John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological methods & research*, 42(3):294–320.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rosanna Cole. 2024. Inter-rater reliability methods in qualitative case study research. *Sociological Methods & Research*, 53(4):1944–1975.

Tobias Daudert. 2020. A web-based collaborative annotation and consolidation tool. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7053–7059.

Dorottya Demszky and Heather Hill. 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 2023)*, pages 528–538.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B 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.

Naihao Deng, Yikai Liu, Mingye Chen, Winston Wu, Siyang Liu, Yulong Chen, Yue Zhang, and Rada Mihalcea. 2023. Ease: An easily-customized annotation system powered by efficiency enhancement mechanisms. *arXiv preprint arXiv:2305.14169*.

Kerry Dhakal. 2022. Nvivo. *Journal of the Medical Library Association: JMLA*, 110(2):270.

Sidney K D'Mello and Art Graesser. 2012. Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5(4):304–317.

Zackary Okun Dunivin. 2025. Scaling hermeneutics: a guide to qualitative coding with llms for reflexive content analysis. *EPJ Data Science*, 14(1):28.

Kathleen M. Eisenhardt. 1989. Building theories from case study research. *Academy of Management Review*, 14(4):532–550. Available at JSTOR: stable 258557.

David Garlan, Vishal Dwivedi, Ivan Ruchkin, and Bradley Schmerl. 2012. Foundations and tools for end-user architecting. In *Large-Scale Complex IT Systems. Development, Operation and Management: 17th Monterey Workshop 2012, Oxford, UK, March 19-21, 2012, Revised Selected Papers 17*, pages 157–182. Springer.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Xudong Hong, Margarita Ryzhova, Daniel Adrian Biondi, and Vera Demberg. 2023. Do large language models and humans have similar behaviors in causal inference with script knowledge? *arXiv preprint arXiv:2311.07311*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Nicholas Hunkins, Sean Kelly, and Sidney D'Mello. 2022. "beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. In *Lak22: 12th international learning analytics and knowledge conference*, pages 230–238.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Emily Jensen, Samuel L. Pugh, and Sidney K. D'Mello. 2021. A deep transfer learning approach to modeling teacher discourse in the classroom. In *LAK21: 11th international learning analytics and knowledge conference*, pages 302–312.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Inc., Thousand Oaks, CA.

Yun Long, Haifeng Luo, and Yu Zhang. 2024. Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, 9(1):60.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. 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.

Neil Mercer and Christine Howe. 2012. Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction*, 1(1):12–21.

Sarah Michaels, Catherine O'Connor, and Lauren B. Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. In *Studies in Philosophy and Education*, volume 27, pages 283–297.

Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated news text. *arXiv preprint arXiv:2308.09067*. Version 3 (Sep 2, 2024).

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Open-source text annotation software.

OpenAI. 2024. Gpt-4o technical report. Accessed: 2025-07-02.

Soya Park, April Yi Wang, Ban Kawas, Q Vera Liao, David Piorkowski, and Marina Danilevsky. 2021. Facilitating knowledge sharing from domain experts to data scientists for building nlp models. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 585–596.

Michael Quinn Patton. 2002. Two decades of developments in qualitative inquiry: A personal, experiential perspective. *Qualitative social work*, 1(3):261–283.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337.

Tal Perry. 2021. Lighttag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27.

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.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662.

Abhijit Suresh, Tamara Sumner, Isabella Huang, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2018. Using deep learning to automatically detect talk moves in teachers' mathematics lessons. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5445–5447. IEEE.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software.

Ludi Wang, Dongze Song, Qiang Cui, Xueqing Chen, Yuanchun Zhou, Wenjuan Cui, and Yi Du. 2025. Autodive+: An adaptive model enhanced multimodal online annotation tool. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2919–2922.

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.

Cynthia Weston, Terry Gandell, Jacinthe Beauchamp, Lynn McAlpine, Carol Wiseman, and Cathy Beauchamp. 2001. Analyzing interview data: The development and evolution of a coding system. *Qualitative sociology*, 24:381–400.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*.

Linxuan Zhao, Dragan Gašević, Zachari Swiecki, Yuheng Li, Jionghao Lin, Lele Sha, Lixiang Yan, Riordan Alfredo, Xinyu Li, and Roberto Martinez-Maldonado. 2024. Towards automated transcribing and coding of embodied teamwork communication through multimodal learning analytics. *British Journal of Educational Technology*, 55(4):1673–1702.