AI-Powered Coding of Elementary Students' Small-Group Discussions about Text

Carla M. Firetto[1], P. Karen Murphy[2], Lin Yan[1], & Yue Tang[2]

[1] Arizona State University; Mary Lou Fulton College for Teaching and Learning Innovation (MLFC); Tempe, AZ, USA
[2] The Pennsylvania State University; College of Education; University Park, PA, USA

Abstract

We present a novel application of an AI-powered approach for coding elementary students' small-group discussions about text. We used AILYZE to identify instances of individual and collective argumentation within a set of 371 transcripts. We gathered evidence of reliability (i.e., via comparability checks with human-produced codes) and criterion validity (i.e., via ground truth checks). There was sufficient agreement between AI-generated and human-produced codes, and initial validity evidence exceeded the established threshold of near-perfect agreement on a small ground truth check. Findings provide evidence that AI may serve to accurately code discussion transcripts in ways that were not previously feasible with only human-produced coding.

1 Introduction

Until recently, educational research examining the use of small-group discussions in preK-20 classrooms has been a resource-demanding area of study. Historically, quantitative analyses have required hand coding by research team members, which comes at significant time and cost expense (Longo, 2019; Murphy et al., 2018; Siiman et al., 2023). Consequently, troves of data often go un- or under-analyzed, yielding the potential loss of innumerable scientific advancements.

Recent developments in artificial intelligence (AI) now provide seemingly unlimited potential regarding automated AI-based discussion coding (Tran et al., 2024; Wang et al., 2024). In the present study, we build on the rapidly advancing work leveraging AI as a tool to code student discussions. Specifically, we present a novel application of an AI tool used to code small-group discussions about text along with the associated evidence of reliability and validity as part of a recent secondary analysis of small-group discussions (Firetto et al., 2025).

1.1 Value of Coding Discussion

There is a large body of research investigating the impact of small-group discussions in preK-20 classrooms and the myriad benefits on various outcome measures (e.g., comprehension, reasoning, transfer, motivation; Bae et al., 2021; Bennett et al., 2010; Murphy et al., 2009). Some of this research examines "after the discussion" or distal outcomes (e.g., class grades or test performance). For example, in our prior research, we found that elementary students engaging in small-group discussions evidenced increases in their written argumentation after discussions about what they read in their language arts class (Firetto et al., 2019; Murphy et al., 2022).

In contrast, other research examines learning based on what occurs "in" or "during" the discussions. For example, in our prior research, we identified indicators of high-level comprehension and tracked their frequency over time (Murphy et al., 2018). Coding and analyzing the discourse directly is particularly beneficial as it may allow for more accurate proximal measures without having to rely on transfer or delayed posttest measures. Researchers have explored a wide variety of coding schemes (Tao & Chen, 2023) and identified a variety of indicators present within the discussion that are associated with high-level comprehension (Soter et al., 2008) and other indicators of academic performance (Howe et al., 2019; Muhonen et al., 2018).

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 125–134

While there is generally a consensus that small-group discussions can benefit students' learning, there are many empirical questions that remain unanswered. For example, little is known about ways to group students: Should groups consist of students with similar or different ability levels? (Murphy et al., 2017); Should students be grouped in single-sex groups or mixed-sex groups? (Bennett et al., 2010). Moreover, there is also much to be learned about whether grade, content, or other factors may serve as moderating variables, ultimately impacting what we know about best practices and the associated recommendations for teachers.

While in-depth, qualitative examinations on smaller samples have contributed important findings toward these ends (e.g., Lobczowski et al., 2020), it is also beneficial for researchers in the field to conduct quantitative examinations derived from large samples (e.g., an experimental study testing multiple different group configurations vs a study with only a treatment and a control). Despite the importance, however, there are massive time and financial costs associated with coding a large corpus of discussion data (Murphy, 2015).

1.2 Leveraging AI to Code Discussion

Given these time and financial costs, researchers have long worked toward finding automated ways to expedite the process of coding. For example, several years ago, we used large language models to derive a series of potential indicators (e.g., complexity, oral expression), which we then compared to comprehension measures (Kosh et al., 2018). This allowed us to identify both word rareness and word diversity as indicators closely associated with students' posttest reading comprehension.

Since then, the ability to leverage AI as a tool to support automated coding processes has grown exponentially (Wang et al., 2024). This shift has meant moving beyond traditional classifiers toward sophisticated, transformer-based systems that track the ebb and flow of classroom talk. For example, in 2021, Song et al. used an artificial neural network-based model to classify the semantic content of classroom dialogue into eight categories. Not surprisingly, their findings indicate that performance (i.e., precision and recall) of the automated coding was better for some categories than others. For example, the prior-known knowledge and analysis categories were high, while other categories, like querying and speculation, were low. The overall F1 score (i.e., a measure of the accuracy of the codes calculated as the harmonic of precision and recall) across all categories was .680.

Advancements in AI are progressing at such a rapid pace that the potential for significant increases in accuracy and speed is growing every day. However, there are three areas, in particular, that need further exploration: (1) additional evidence demonstrating reliability and validity of automated codes along with comparisons to human coding; (2) transparency regarding the coding in ways that ensure model decisions are interpretable to teachers and researchers (i.e., *explainability*); and (3) closing the gap in AI-and-discourse research, whereby studies prioritize model building over *in situ* evaluation (Wang et al., 2024). Together, these issues underscore the need for further study and exploration.

1.3 The Present Study

Over the past year and a half, we conducted a secondary analysis of small-group discussions based on a large set of previously uncoded video-recorded small-group discussions collected as part of a large federally funded grant (R305A130031). We employed an AI-powered coding approach that allowed us to examine changes in students' individual and collective argumentation over time while also investigating the roles of genre and grade-level (Firetto et al., 2025).

Specific to the aims of AIMEcon (i.e., the theme "validity and reliability of AI-driven automated scoring systems"), the present study extends our previous work by examining the comparability of AI- and human-coded outcomes as well as the ways in which AI-powered coding can be rigorously employed. We explored two primary RQs:

RQ1: Are codes produced by AILYZE roughly comparable to those previously produced by humans? Does Cohen's Kappa agreement between AI-generated codes and human-produced codes meet or exceed .60 (i.e., *substantial* agreement)?

RQ2: Are codes produced by AILYZE accurate, based on a ground truth check? Does Cohen's Kappa agreement between AI-generated codes and human verifications meet or exceed .80 (i.e., *near-perfect* agreement)?

2 Method

2.1 Sample

The sample consisted of 371 transcripts of small-group discussions (i.e., typically 4-6 students per group). The discussions were conducted in fourth- and fifth-grade classrooms and collected over an entire school year. 3PlayMedia produced the transcriptions from video recordings using professional human transcribers (see Murphy, 2025). In accordance with our IRB protocol, research team members cleaned the transcripts to remove identifying information before entering the files into AILYZE (see Appendix A for detailed specifications).

2.2 Codes

We focused on identifying instances of two specific discourse indicators of high-level comprehension (i.e., individual and collective argumentation). *Individual argumentation* was intended to capture instances where a specific student produced an extended response that included multiple pieces of argumentation (e.g., a claim supported by reasoning and evidence). For example, a student explained, "I'd feel brave because, if I were Sahar, I would be going past the limit where I was supposed to be swimming. And I would be kind of a hero for saving that tiger." This example illustrates individual argumentation as it included a specific claim about how the student believed they would feel if they were the main character (i.e., brave), along with two pieces of support for that claim (i.e., risk-taking by going beyond their swimming boundary and rescuing a tiger). Individual argumentation codes are informed by the notion of elaborated explanation (Chinn et al., 2000; Webb, 1991), which is a well-established discourse indicator of high-level comprehension.

Collective argumentation, on the other hand, represented episodes of talk where two or more students co-constructed understanding together. Importantly, our coding definition required the inclusion of an element of disagreement (e.g., a challenge or counterargument). For example, the discussion excerpt presented in Table 1 represents collective argumentation.

This example illustrates collective argumentation as it included multiple turns of students exploring the idea about whether the story was realistic or not, specifically the notion of whether "saving a polar bear" is something that one could realistically do, particularly given the massive size of adult polar bears. Collective argumentation codes are informed by the concept of exploratory talk (Mercer, 1995, 2000), another well-established discourse indicator of high-level comprehension.

Student A	Anything's possible, especially something that is realistic. [referring to a previous statement about whether the story they read was something that could happen in real-life]
Student B	Except trying to save a polar bear. Might just be scared.
Student A	You could save polar bears.
Student B	Well, yeah, you could. Except, by the way, it is heavy because its really heavy. They almost weigh, like, thousand millions of pounds. [inaudible/interposing voices] Still, it's heavy.
Student A	They actually weigh, like 1,000 pounds.
Student B	Still, its heavy and bigger. You can get crushed.
Student C	Not a baby one.

Table 1: Talk Excerpt of Collective Argumentation

2.2.1 Previous Approach to Coding: Human-Produced Codes

Before new research assistants on our team code independently, they begin with an orientation to coding (e.g., reading the coding manual and related standard operating procedures), learn about the video recording software, and receive extensive mentoring with an experienced coach. Over the past decade, we have documented that it takes new human coders approximately 40 hours of coding training and practice to become relatively proficient at coding the recordings of small-group classroom discussion using our coding manual. Moreover, even after they have demonstrated proficiency, research assistants continue to engage in regular fidelity checks. Thus, as coders engage in the coding over time, 20% of the recordings are independently coded by a second research assistant, the codes are compared, and the two research assistants justify to each other why they

coded or didn't code a specific event where there was a point of disagreement and then come to an understanding about which is the best fit. While time consuming, this procedure helps to maintain fidelity to the codebook. This is also due, in part, because one of the guiding principles of the codebook is to maintain low levels of inference (e.g., not to assume a student's intention or meaning), thus these fidelity checks also serve as an accountability check toward this standard.

2.2.2 Novel Approach to Coding: AI-generated Codes

In line with existing Human-in-the-Loop approaches to discourse analysis (e.g., Cohn et al., 2024), we leveraged AILYZE to produce AI-generated codes in a way that augmented, not replaced, expert judgment. In our case, we employed an approach in which the research team defined the codebook, designed the prompts, and decided the acceptance criteria prior to large-scale transcript coding for individual and collective argumentation.

Our initial plan involved deriving training examples from a sample of previously human-coded discussions to fine-tune the AI model to increase the coding accuracy. To do this, we transferred the human-produced codes from the video recording coding software onto the transcribed text documents for a sample of discussions that had been coded by two research assistants. However, the codes derived from the video recordings did not always translate directly and accurately to the transcripts. For example, in some cases, during the video recording, it was clear to observers that a single student articulated individual argumentation within a given turn, yet on the transcription, it might appear that this turn was interrupted (i.e., a student speaking over another and cutting one turn into multiple turns). In addition, the human coders in our lab and the humans who transcribed the recordings at 3PlayMedia may have had differences in what they heard and understood during the discussion, influencing how a word or phrase was interpreted or dismissed as inaudible. Ultimately, we decided not to include human-produced codes in the training of the AI model, and instead we used them to conduct a comparability check (i.e., RQ1).

Across multiple iterations, we revised a prompt informed by the definitions and coding criteria established in the Quality Talk coding manual (Murphy et al., 2017). We reviewed the AI-generated codes and corresponding justifications to refine the prompt, adding additional details as needed (e.g., the role of the teacher) and fixing data mapping issues (e.g., rows without dialogue).

Once the first and second authors independently agreed that we had developed a prompt that led to sufficiently accurate AI-generated codes, we conducted: (a) a comparability check, in which AI-generated codes were compared with previous human-produced codes from the video recordings, and (b) a ground truth check, in which the first two authors (each with hundreds of hours of discourse coding experience) collectively manually coded two transcripts to serve as the reference standard. Because both checks exceeded our pre-established thresholds (see results below), the AILYZE model was then applied to the deployment phase, coding the full set of transcripts (see Appendix B and Appendix C).

3 Results

3.1 RQ1

For the first research question, we examined the extent to which AI-generated codes were comparable to those previously produced by humans. We compared the AI-generated codes to the human-produced codes using 37 transcript excerpts (i.e., 10% of the total number) containing 3,249 turns. Due to the aforementioned difference in modality (i.e., transcript vs. video coding), we set our Cohen's Kappa threshold at .60, representing at least *substantial* agreement. Both codes exceeded this threshold: individual (Cohen's Kappa = 0.735, SD = 0.022, 95% CI [0.691, 0.775]) and collective argumentation (Cohen's Kappa = 0.849, SD = 0.014, 95% CI [0.823, 0.875]). Overall, there was sufficient consistency between the AI-generated codes from the transcripts and previously produced human codes from the video recordings.

To better illustrate the impact that modality may have had on coding, we identified an example from one of the discussions where there was disagreement between the AI-generated code and the human-produced code. The discussion excerpt in Table 2 begins after a statement made about there being lots of things to do outside.

Student C	I agree with Student B's idea because I would [INAUDIBLE] my house a lot. So I usually go outside. I have a trampoline so I can jump on that. But I go outside, and I pretend that I'm going to teach her something. I have magical powers and [INAUDIBLE].
Teacher	[Chuckles]
Student C	Because there's like—
Teacher	[Chuckles] I pretend I'm a teacher with magical powers too sometimes.
Student C	Yeah. Because we have a wooden hat. I get a stick and then [INAUDIBLE] or something.

Table 2: Talk Excerpt Illustrating AI/Human Disagreement

The first turn by Student C was identified as an instance of individual argumentation by AILYZE with the justification: *The claim is 'I agree with Student B's idea' supported by the reasons 'I usually go outside' and 'I pretend that I'm going to teach her something'. This provides a claim + multiple reasons + personal experience as evidence.* It is relatively clear from both the transcript and the AILYZE justification that this turn meets the criteria for an individual elaboration code. However, this turn was not coded by the research assistants. While there are many possible reasons why the human coders did not identify this turn as an instance of individual argumentation, the larger transcribed excerpt illustrates two possible explanations: (a) The teacher's chuckles and verbal/non-verbal input may have interrupted or influenced the student's talk as the human coders watched the video, which does not seem to be the case based on the way it was transcribed; (b) The human coders may have understood more or less of the words than the professional transcriber, who already noted "[INAUDIBLE]" in several places. This could have influenced the research assistants' decision to identify this turn as an instance of individual argumentation (e.g., hearing words that may have changed the meaning, aiming for a low inference interpretation of what they actually could hear).

3.2 RQ2

For the second research question, we aimed to gather evidence regarding whether the AI-generated codes were accurate based on a small ground truth check, where we selected two of the transcripts with both AI-generated and human-produced codes and then verified the accuracy of the codes at each turn (i.e., $n = 144$ transcript turns). Because we conducted the ground truth check using the text transcripts, we aimed for *near-perfect* Cohen's Kappa (i.e., at least .80) for the AI-generated codes. Both codes exceeded this threshold: individual (Cohen's Kappa = 1.00, SD = 0, 95% CI [1.00, 1.00]) and collective argumentation (Cohen's Kappa = 0.959, SD = 0.017, 95% CI [0.926, 0.991]). The AI-generated codes exceeded the pre-established threshold.

Given that the ground truth check was performed on transcripts that also had the human-produced codes (i.e., transferred from the video recordings), we also conducted an exploratory calculation of the Cohen's Kappa agreement for the human-produced codes: individual (Cohen's Kappa = 0.773, SD = 0.084, 95% CI [0.608, 0.937]) and collective argumentation (Cohen's Kappa = 0.573, SD = 0.041, 95% CI [0.493, 0.654]).

While it is important to underscore again the differences in modality (i.e., coding the transcripts vs coding the video recordings) as well as the relatively small sample size (i.e., two transcripts; $n = 144$ turns), it is noteworthy that both of the Cohen's Kappa values were higher for the ground truth-to-AI-generated codes than they were for ground truth-to-human-produced codes and that there was no overlap in the confidence intervals.

4 Conclusions

Human coding of qualitative data can be extensively resource-intensive (Longo, 2019; Siiman et al., 2023). AI-powered coding can decrease the resources needed to conduct such research and allow for scientific advancements that may not have been previously feasible (Feuston & Brubaker, 2021; Lixandru, 2024; Siiman et al., 2023; Tran et al., 2024). Our findings suggest that AI can be used to code discourse transcripts consistent with human coders when prompts and rules can be derived from established codes and evidence-based manuals (e.g., Murphy et al., 2017).

Notably, our AI coding was completed in a drastically less time than human coding would permit. As a point of reference, experienced coders require about an hour to code one small-group discussion. For this sample, that would have required roughly 371 hours of coding for the research assistants, a load which is typically split (i.e., about 185 hours each) between two coders. In addition, each coder would spend an additional 37 hours of coding the discussions for the fidelity check (i.e., 20%, as described above) and another 37 hours meeting with each other to discuss the instances of agreement/disagreement. Taken together, each coder would need to devote about 259 hours to coding the discussions, assuming they could code continuously. In our funded projects, we have hired undergraduate and graduate research assistants to code the discussions. Typically, graduate research assistants work 20 hours per week and attend classes. As a result, it would take the research assistants nearly an entire semester (i.e., 13 weeks) to complete the coding; however, in our experience, even expert humans cannot code accurately over long periods of time. As such, a more realistic estimate is that it would take two research assistants the better part of an academic year to code this many hours of video.

In contrast, AILYZE processed all 371 transcripts within ~12 hours (i.e., 2 minutes per transcript), clearly illustrating the potential to save time. Importantly, however, human time is still required to develop the prompts and check the AI over time, just as it is required as part of the process for developing, mentoring, and supporting research assistants when doing human-produced coding. Moreover, beyond these efficiency gains, reallocating research assistants' time to more enriching activities could help to move the field forward via increased productivity and dissemination possibilities (e.g., assisting in writing manuscripts, interpreting the data) and by better preparing them for their future research and career endeavors. Beyond the time costs, however, financial costs, environmental costs, energy costs and other costs need to be carefully considered and weighed.

Leveraging AI also has the potential to enhance coding consistency by reducing sources of variability that are common with human coders. As mentioned, human coders can be affected by limitations such as fatigue, overload, or selective attention when working with lengthy qualitative texts (Miles & Huberman, 1994). AI systems, by contrast, apply the same coding criteria uniformly across large corpora without a decline in performance over time. Although prior work has noted that AI-only coding can yield limited reliability (e.g., Prescott et al., 2024), we found that once the AI prompt was refined with human review, it achieved strong alignment with human codes. These findings highlight that while AI contributes efficiency and consistency, human oversight remains essential for guiding the framework, validating outputs, and ensuring methodological rigor.

A further benefit of using AILYZE was its capacity to generate explanations and justifications for coding decisions. This provided transparency that strengthened our human-in-the-loop process. Both in the validation and final coding review processes, we could monitor the rationales for AI-generated codes, which enabled us to identify points of alignment and divergence with human reasoning. This transparency underscores how AI can complement, rather than replace, expert judgment.

Importantly, within the context of our analysis, it is critical to note that the lower rates of agreement for the ground truth analysis and the human-produced codes do *not* negatively reflect on the quality of the human-produced coding. Rather, they highlight the role of modality differences. The human coders were coding with higher accuracy with video and audio, while AI coded based on the text transcripts. In essence, text "stands still." The reliability check, therefore, was to evaluate whether, despite these modality differences, the overall patterns of coding remained comparable across human and AI coders. Future research must take into consideration the potential impact of modality on the codes (e.g., transcripts vs. videos).

Our approach prioritized student privacy, data governance, and responsible use. In accordance with our IRB protocol, all transcripts were de-identified by the research team before any AI processing. We selected AILYZE in part because of its security features, auditability, and policy not to train on user data. To mitigate automation bias and address opacity, we used the aforementioned human-in-the-loop workflow, comparability, and ground truth checks prior to large-scale deployment as well as careful attention to known modality differences between transcripts and video recordings that can shape meaning. Collectively,

these measures aim to reduce risks while ensuring transparent, auditable, and pedagogically responsible use of classroom discourse data.

4.1 Future Research

In the present study, we were able to document an AI-powered coding process that took less time than traditional human coding. However, future research in this area should also consider other important variables beyond time, including costs such as environmental impacts, electricity and/or water usage (Kandemir, 2025), and conduct a true cost–benefit analysis.

During our oversight process, we noted that students' individual and collective argumentation represented a variety of quality (e.g., some students made sophisticated arguments with counter-arguments, rebuttals that extended over a long period of time, while others were more simplistic and succinct). Moving forward, it is critical to understand more about the quality of students' responses, beyond just coding the presence or absence of argumentation within the discussions. While such coding may not be feasibly possible with human coding, we think that it is possible to extend the procedures we employed herein to move beyond binary codes (i.e., presence or absence of a code for a given turn) to develop an AI-generated *quality* score that can capture characteristics such as accuracy, depth, and length.

Finally, now that we have established an automated coding that meets the requisite criteria we established, we can begin to explore how other aspects can be automated. For example, Li et al., 2025 found that AI-generated feedback about classroom discussions was useful for teachers, and thus, coding may also be used to support coaching.

Acknowledgments

We acknowledge the contributions of our collaborators related to the larger project that this study was drawn from: Emily Starrett, Emilee A. Herman, and Jeffrey A. Greene. We also appreciate the support of James Goh for assisting us during the AI-powered coding procedures and for providing detailed descriptions of AILYZE (see Appendix A).

Funding

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130031 to the Pennsylvania State University. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not represent the views of the Institute or the U.S. Department of Education. Additional support was also provided by MLFC internal grant funding at Arizona State University and McMichael Professorship in the School of Education at the University of North Carolina at Chapel Hill. This material is based upon work supported by the National Science Foundation under Award No. (1912415). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest

The authors report no known conflict of interest. The authors purchased use of AILYZE for the purpose of this project and have no financial relationship with AILYZE.

References

AILYZE, Inc. (n.d.). *AILYZE*. https://www.ailyze.com/

Bae, C. L., Mills, D. C., Zhang, F., Sealy, M., Cabrera, L., & Sea, M. (2021). A systematic review of science discourse in K–12 urban classrooms in the United States: Accounting for individual, collective, and contextual factors. *Review of Educational Research*, *91*(6), 831-877. https://doi.org/10.3102/00346543211042415

Bennett, J., Hogarth, S., Lubben, F., Campbell, B., & Robinson, A. (2010). Talking science: The research evidence on the use of small group discussions in science teaching. *International Journal of Science Education*, *32*(1), 69-95. https://doi.org/10.1080/09500690802713507

Chinn, C. A., O'Donnell, A. M., & Jinks, T. S. (2000). The structure of discourse in collaborative learning. *The Journal of Experimental Education*, 69(1), 77–97. https://doi.org/10.1080/00220970009600650

Cohn, C., Snyder, C., Montenegro, J., & Biswas, G. (2024). Towards a human-in-the-loop LLM approach to collaborative discourse analysis. In *International Conference on Artificial Intelligence in Education* (pp. 11–19). Springer Nature Switzerland.

Feuston, J. L., & Brubaker, J. R. (2021). Putting tools in their place: The role of time and perspective in human-AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. https://dl.acm.org/doi/pdf/10.1145/3479856

Firetto, C. M., Murphy, P. K., Greene, J. A., Li, M., Wei, L., Montalbano, C., Hendrick, B., & Croninger, R. M. V. (2019). Bolstering students' written argumentation by refining an effective discourse intervention: Negotiating the fine line between flexibility and fidelity. *Instructional Science,* 47, 181–214. https://doi.org/10.1007/s11251-018-9477-x

Firetto, C. M., Murphy, P. K., Starrett, E., Herman, E. A., Greene, J. A., Tang, Y., & Yan, L. (2025). Investigating grade-level and text genre effects in Quality Talk discussions: An AI-powered discourse analysis of upper primary students' high-level comprehension. *Learning and Instruction*, 100, 102208. https://doi.org/10.1016/j.learninstruc.2025.102208

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Kandemir, M. (2025, April 8). *Why AI uses so much energy-and what we can do about it*. Institute of Energy and the Environment. https://iee.psu.edu/news/blog/why-ai-uses-so-much-energy-and-what-we-can-do-about-it

Kosh, A. E., Greene, J. A., Murphy, P. K., Burdick, H., Firetto, C. M., & Elmore, J. (2018). Automated scoring of students' small-group discussions to assess reading ability. *Educational Measurement: Issues and Practice,* 37(2), 20–34. https://doi.org/10.1111/emip.12174

Li, X., Han, G., Fang, B., & He, J. (2025). Advancing the in-class dialogic quality: Developing an artificial intelligence-supported framework for classroom dialogue analysis. *Asia-Pacific Education Researcher,* 34, 495–509. https://doi.org/10.1007/s40299-024-00872-z

Lixandru, D. (2024). The use of artificial intelligence for qualitative data analysis: ChatGPT. *Informatica Economica*, *28*(1). https://doi.org/10.24818/issn14531305/28.1.2024.05

Lobczowski, N. G., Allen, E. M., Firetto, C. M., Greene, J. A., & Murphy, P. K. (2020). An exploration of social regulation of learning during scientific argumentation discourse. *Contemporary Educational Psychology,* 63, https://doi.org/10.1016/j.cedpsych.2020.101925

Longo, L. (2019). Empowering qualitative research methods in education with artificial intelligence. In *World Conference on Qualitative Research* (pp. 1–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-31787-4_1

Mercer, N. (1995). The guided construction of knowledge: Talk amongst teachers and learners. Multilingual Matters.

Mercer, N. (2000). *Words and minds: How we use language to think together*. Routledge. https://doi.org/10.4324/9780203464984

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.

Muhonen, H., Pakarinen, E., Poikkeus, A. M., Lerkkanen, M. K., & Rasku-Puttonen, H. (2018). Quality of educational dialogue and association with students' academic performance. *Learning and Instruction*, *55*, 67-79. https://doi.org/10.1016/j.learninstruc.2017.09.007

Murphy, P. (2025). Investigating grade-level and genre effects in Quality Talk discussions: An AI-powered discourse analysis of upper primary students' high-level comprehension. *Databrary*. https://databrary.org/volume/1858

Murphy, P. K. (2015). Mooring points and touchstones along the road to school-based interventions—An introduction. *Contemporary Educational Psychology, 40*, 1-4. https://doi.org/10.1016/j.cedpsych.2014.10.003

Murphy, P. K., Firetto, C. M., Greene, J. A., & Butler, A. M. (2017). *Analyzing the talk in Quality Talk discussions: A coding manual*. http://doi.org/10.18113/S1XW64

Murphy, P. K., Greene, J. A., Firetto, C. M., Croninger, R. M., Duke, R. F., Li, M., & Lobczowski, N. G. (2022). Examining the effects of Quality Talk discussions on 4th- and 5th-grade students' high-level comprehension of text. *Contemporary Educational Psychology,* 71, https://doi.org/10.1016/j.cedpsych.2022.102099

Murphy, P. K., Greene, J. A., Firetto, C. M., Li, M., Lobczowski, N. G., Duke, R. F., & Croninger, R. M. (2017). Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension. *Contemporary Educational Psychology*, *51*, 336-355. https://doi.org/10.1016/j.cedpsych.2017.09.003

Murphy, P. K., Greene, J. A., Firetto, C. M., Hendrick, B., Li, M., Montalbano, C., & Wei, L. (2018). Quality Talk: Developing students' discourse to promote high-level comprehension. *American Educational Research Journal,* 55(5), 1113–1160. https://doi.org/10.3102/0002831218771303

Murphy, P. K., Wilkinson, I. A., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology*, 101(3), 740. https://doi.org/10.1037/a0015576

Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*, e54482. https://doi.org/10.2196/54482

Siiman, L. A., Rannastu-Avalos, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023, August). Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In *International Conference on Innovative Technologies and Learning* (pp. 87-96). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40113-8_9

Song, Y., Lei, S., Hao, T., Lan, Z., & Ding, Y. (2021). Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research,* 59(3), 496–521. https://doi.org/10.1177/0735633120968554.

Soter A. O., Wilkinson I. A. G., Murphy P. K., Rudge L., Reninger K., Edwards M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47, 372–391.

Tao, Y., & Chen, G. (2023). Coding schemes and analytic indicators for dialogic teaching: A systematic review of the literature. *Learning, Culture and Social Interaction*, *39*. https://doi.org/10.1016/j.lcsi.2023.100702

Tran, N., Pierce, B., Litman, D., Correnti, R., & Matsumura, L.C. (2024). Analyzing large language models for classroom discussion assessment. arXiv. https://doi.org/10.48550/arXiv.2406.08680

Wang, D., Tao, Y., & Chen, G. (2024). Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, *123*. https://doi.org/10.1016/j.ijer.2023.102275

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education*, *22*(5), 366-389.

5 Appendices

5.1 Appendix A

AILYZE Specifications: Reported by AILYZE

AILYZE's LLMs (i.e., a mix of Grok-1, Mistral 8x22B, and Phi-2.5-MoE) are trained on curated, high-quality open corpora commonly used in multilingual and scholarly modeling, such as the UN Parallel Corpus (multilingual proceedings and debates), ParlaMint (TEI-standard legislative proceedings with speaker-linked context), and S2ORC (millions of scholarly articles with citations and structure). Importantly for the context of the present study, it is also trained on education-specific transcripts, such as TalkBank's ClassBank (curated classroom discourse collections), the NCTE Elementary Math Classroom Transcripts (1,660 lessons from 4th–5th grade), and the Teacher–Student Chatroom Corpus (one-to-one teacher–learner lessons). This mix supports discourse-focused tasks, including segment-level coding with grounded justifications.

In addition, AILYZE complies with measures outlined in the HECVAT (Higher Education Community Vendor Assessment Toolkit), which is a standardized framework developed by higher education organizations to assess data and AI risks associated with technology services. AILYZE does not train on user data and all project data, prompts, and outputs are encrypted, access-controlled, and exportable for archiving. All runs are also versioned so that the same codebook, same engine version and same transcripts yield identical results, ensuring full reproducibility.

AILYZE's deterministic inference setting was used, which locked the codebook and engine version for the entire run. This ensures that if future researchers re-run the same transcripts with the same project configuration, they will obtain identical labels and justifications, supporting fully reproducible analyses.

5.2 Appendix B

Individual Argumentation Prompt:
"Code the interview transcript to identify all instances of elaborated explanations. Elaborated explanations are instances in which students explain their thinking in fairly coherent form to others. They occur in a single turn where a student explains how he or she arrived at a conclusion or idea by giving a step-by-step description or detailed account of how the conclusion or idea was reached or how a problem might be resolved. They are elaborated descriptions of how things work, why some things are the way they are, or how they should be thought about. They include details of how to think about an issue and justification or rationale for thinking that way. Elaborated explanations relate to the quality of explanations given by an individual student, not a collective of students, and not the teacher. They can take various forms including: claim + 2 or more independent reasons, claim + 2 or more conjunctive reasons, claim + 2 or more causally connected reasons, claim + reason(s) + evidence, claim + reason(s) + warrant, or claim + evidence + evidence. Elaborated explanations must include all components within a single "turn." The components begin at the start of the claim and continue through the end of the speaker's turn, unless the topic shifts away from that claim. A claim may be implied in verbal discourse when it immediately follows a question, but the response must directly respond to or follow from a question within the same question event. Elaborated explanations can only occur within authentic question events; responses to test questions cannot be coded as elaborated explanations."

5.3 Appendix C

Collective Argumentation Prompt:
"Code the interview transcript to identify all instances of exploratory talk. Exploratory talk occurs when students share, evaluate, and build knowledge over at least three turns. It is talk in which partners engage critically but constructively with each other's ideas where relevant information is offered for joint consideration. Proposals may be challenged and counter-challenged but, if so, reasons are given and alternatives are offered. Agreement is sought as a basis for joint progress, with knowledge made publicly accountable and reasoning visible in the talk. It embodies a kind of 'co-reasoning,' with speakers following ground rules which help them to share knowledge, evaluate evidence, and consider options in a reasonable and equitable way. The key component of exploratory talk is the element of challenge, with only one challenge statement necessary for an episode to be classified as exploratory talk. Exploratory talk episodes consist of instances where students co-construct understanding over at least three consecutive, uninterrupted turns about the same topic. Exploratory talk is characterized by students actively constructing knowledge with students primarily interacting with and talking to each other. Episodes end when the topic shifts, someone asks a different question, a statement is made that deviates from the trajectory, or the students arrive at consensus. A student must initiate the challenge for talk to be considered exploratory. The teacher can be present but is not influencing the discourse or episode of talk. Exploratory talk in essence is a way of using language to think collectively—to 'interthink.'"