Human and LLM-based Assessment of Teaching Acts in Expert-led Explanatory Dialogues

Anonymous ACL submission

Abstract

Didactics research explores what instructional strategies yield the best learning outcomes in expert-led explanatory dialogues on complex concepts. In the paper at hand, we address this question by annotating a dataset of dialogues that foster scientific understanding, categorized across five levels of the explainee's knowledge. Our extended dataset, ReWIRED, features span-level annotations for teaching-related explanatory acts. Furthermore, we assess language models of varying sizes on their ability to label teaching acts, uncovering that fine-tuning is necessary for modeling the task, especially GPT-4o-mini with structured prediction profits from that. Finally, we leverage and extend a set of quality metrics for instructional explanations, by involving annotators to estimate the relevance and impact of each metric across the five knowledge levels. We then apply the metrics to our newly annotated dialogues and expand it into a prompt-based framework, enhancing its applicability and scope. Our findings reveal a strong alignment between the quality metrics and the knowledge levels, with expert explanations in our dataset frequently reflecting established best teaching practices.

1 Introduction

Large language models (LLMs) have notably advanced the integration of artificial intelligence (AI) into human-computer interaction and its application to specific domains. This development impacts the field of explainable artificial intelligence (XAI), particularly the generation of natural language explanations. For explanations, XAI can draw on existing research from various disciplines, including philosophy, cognitive psychology, and social psychology (Miller, 2019). Insights from the field of education can provide valuable guidance for developing explanatory dialogues between *explainers* and *explainees*. Conversely, AI has the potential to contribute significantly to educational practices.

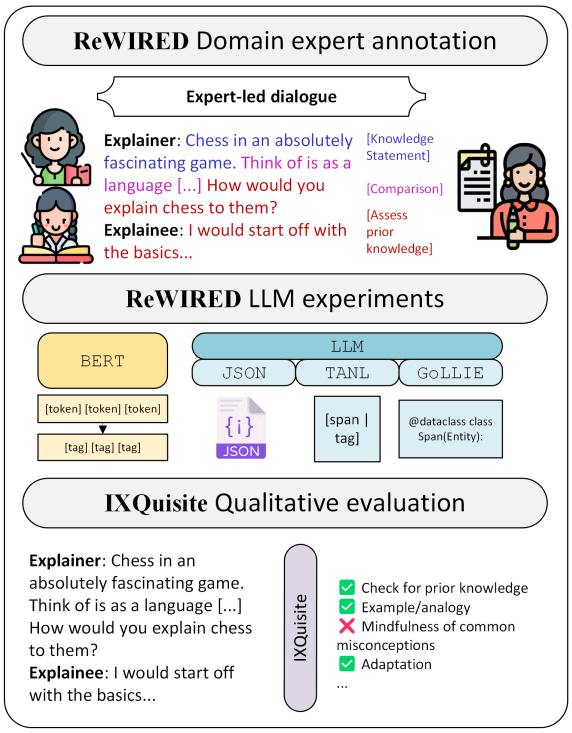


Figure 1: After acquiring span-level dialogue annotations from education domain experts, we conduct experiments to evaluate LLMs' performance in predicting teaching act-related labels across various output formats. Human and LLM-based qualitative evaluation at each level provides deeper insights into model capabilities.

The perception of AI's role in education is diverse and depends on the educators' familiarity with it (Kasinidou et al., 2024). A growing body of research explores methodologies for effectively incorporating AI technologies into educational frameworks for various tasks, including the quality assessment of explainees' answers (Carpenter et al., 2024) as well as their cognitive engagement (McClure et al., 2024). In contrast, Feldhus et al. (2024) concentrated on the explainer side of the dialogue, providing span labels of teaching acts in dialogues between an expert in a given field and different explainees (ranging from laypersons to colleagues).

In this work, we study the quality of the explanations provided by explainers in dialogues. Building on the annotation scheme of Feldhus et al. (2024), we annotate the WIRED dataset presented by Wachsmuth and Alshomary (2022) with the help of teaching expert annotators. We double the size of the original dataset by adding 65 dialogues on 13 new topics released later (§3). We argue that span-labeling is preferable to classification for our dataset, because it enables the precise annotation of instructional segments within dialogues, preserving contextual dependencies and allowing for overlapping or nested teaching strategies.

Building on this newly annotated dataset, we evaluate the performance of language models such as `BERT` (Devlin et al., 2019) and LLMs including `GPT-4o` (OpenAI, 2023) and `Gemini` (Reid et al., 2024) in predicting teaching acts (§4). The results reveal that LLM performance on such a span-labeling task is highly sensitive to requested output formats: We find that structured prediction with JSON output from LLMs poses challenges for post-processing, but they can be mitigated by few-shot demonstrations, improving consistency and performance. However, alternative output-structuring approaches based on inline tagging and code generation (Paolini et al., 2021; Sainz et al., 2024) achieve significantly better outcomes, particularly for complex teaching acts. Notably, when fine-tuned on a subset of the data, BERT outperforms the LLMs across most classes, and a `GPT-4o-mini` fine-tuned on inline tagging excels across the board, highlighting the importance of controlled setups for span-labeling applications.

To assess the quality of expert-led explanations, we refine a set of automated quality metrics introduced in literature with human validation and extension to include LLMs "as a judge" in the evaluation process. The metrics evaluate characteristics that enhance the quality of instructional explanations when present in a dialogue (§5). The metrics fall into two categories: *functional*, which assess the presence of various teaching acts within a dialogue, and *form-based*, which analyze linguistic features such as syntactic and lexical complexity.

We validate our test suite with expert annotators who assess the presence and contribution of all quality metrics within each dialogue. Additionally, we extend the existing metrics with prompt-based metrics, following the methodology of Rooein et al. (2024). Our findings show that metrics that were previously difficult to capture in an automated way align well with the five explainee knowledge levels when using our new prompt-based variants.

Altogether, our four main contributions are:

- The extension of an existing dataset by further explanatory dialogues and by expert span-level annotations of teaching acts;
- The empirical evaluation of the ability of language models to label the teaching acts within the newly proposed ReWIRED dataset;
- The validation of various metrics for dialogical explanation quality with annotators to assess their relevance and impact across different levels of expertise;
- The employment of a prompt-based evaluation framework which broadens the scope and usability of the instructional quality metrics, facilitating their use in diverse scenarios.

An overview of our contributions and workflow can be seen in Figure 1.

2 Background and related work

Instructional explanations are intended to transfer knowledge by introducing a new cognitive framework for understanding a concept or performing a task, bridging the gap between a knowledgeable individual and someone lacking that understanding. In science education, such explanations are considered both a fundamental activity and a goal of scientific practice, aimed at systematically addressing "how" and "why" questions (Kulgemeyer, 2018). The authors highlight the separation of two interpretations for the term *explanation*: One is an explanation seen as activity, whose goal is to "engender understanding" between an explanation holder and an explainee; the other is a more philosophical understanding explanation, as that which connects *explanans* and *explanandum* (Zhu and Rudzicz, 2023). Although most studies concerning explainability have focused on the latter (Miller, 2019), we see the former as the most important definition, as it directly relates to the contextual setting of explanation. Among explanation types, we concern ourselves with *instructional* explanations, a concept from didactics that means "to convey a procedure or model of how to interpret the world between two interlocutors" (Kulgemeyer, 2018).

Teaching models are structured approaches designed to guide educators in planning lessons more effectively, aligning them with psychological learning principles to enhance student outcomes. They are related but different from *learning models*,

which in turn seek to explain how learning happens in the mind of the students. While there have been attempts at unifying multiple teaching and learning models (explaining how learning happens in the mind of the students) (Oser and Baeriswyl, 2002), many remain sceptical about the feasibility (Allensworth et al., 2008). Boston (2012) abstracted the differences and used broad definitions of the processes, leading to positive outcomes but failing to evaluate low-level, dialogical components of teaching. Teaching processes are here represented in the form of teaching acts (Table 1, Table 4) and as explanation quality metrics (Table 5). In the former case, we investigate if language models can capture the distinctions, while in the latter, we conduct an analysis of their correlation to the knowledge levels of the explainees.

Existing dialogue datasets such as CIMA (Stasaski et al., 2020), TSCC-2 (Caines et al., 2022), and NCTE (Demszky and Hill, 2023), focus on surface-level interaction between teachers and students. Our work is closest to Wachsmuth and Alshomary (2022), who annotated a conversation corpus with dialogue acts and explanation moves. More recently, Alshomary et al. (2024) introduced a corpus of explanatory dialogues sourced from Reddit. Their annotation schema closely mirrors that of Wachsmuth and Alshomary (2022), providing a comparative analysis in terms of dialogue and explanation moves, as well as dialogue act flow, with the WIRED dataset.

AI/LLMs in education Recent advancements in leveraging large language models (LLMs) have shown significant potential in educational contexts. Carpenter et al. (2024) investigate using LLMs to evaluate the correctness of explanations provided by undergraduate computer science students. McClure et al. (2024) explore LLMs as tools for classifying cognitive engagement levels. Wang et al. (2024) and Jurenka et al. (2024) introduce collaborative human-AI systems that provide educators with expert-like guidance during tutoring sessions, facilitating the identification and reinforcement of effective pedagogical strategies while discouraging less effective ones.

Evaluation of dialogues and instructional explanations While automatic metrics have been proposed for the quality of discourse and explanation (McNamara et al., 2014; Demszky et al., 2021; Schuff et al., 2023), research has recently been focussing on LLM-based evaluation. Mehri and Eskénazi (2020) propose reference-free quality metrics to evaluate dialogues automatically on both turn and dialogue levels. Rooein et al. (2024) assess difficulty and readability of texts in various levels with *static* (automated) and *prompt-based* metrics. Xu et al. (2024) assess high-level instruction quality using LLMs and stress that these perform on par with human raters for straightforward, discrete variables requiring little inference, but they struggled with analyzing complex teaching practices.

Teaching Act	T. Mdl.
T01: *Assess Prior Knowledge* Checking what the student knows before starting a lesson	CB, UT
T02: *Lesson Proposal* Proposing the steps that will be taken during the lesson	UT
T03: *Active Experience* Providing the student with puzzle/question to explore; (Student:) Interacting with a mental concept	CB, UT
T04: *Reflection* Finding gaps in knowledge or inconsistencies; Asking questions about the experience or concept	PS
T05: *Knowledge Statement* Stating the concept(s) being taught via rules or facts	PS
T06: *Comparison* Considering similarities and differences between the main concept and other related topics or facts	UT
T07: *Generalization* Exploring how the concept applies to new scenarios, experiences and situations outside of the lesson topic	CB, PS
T08: *Test Understanding* Finding out if the concept previously established was received correctly and is properly understood	CB
T09: *Engagement Management* Maintaining the classroom context to facilitate effective teaching, creating rapport between teacher and student	

Table 1: Teaching acts in the ReWIRED dataset (with descriptions and their connection to a teaching model from didactics: Teaching as problem solving (**PS**), teaching as concept building (**CB**) (Krabbe et al., 2015), and unified teaching choreographies (**UT**) (Oser and Baeriswyl, 2002).

3 The ReWIRED dataset

In this section, we present our ReWIRED dataset, an extension of an existing corpus that we propose to study the instructional strategies in explaining dialogues. In the following, we detail the source data as well as the annotation scheme and process.

3.1 Source data: Explanation dialogues

We build on the WIRED corpus (Wachsmuth and Alshomary, 2022), which consists of instructional explanatory dialogues retrieved from the 5-Levels video series[1]. The edited video clips demonstrate how an expert explains (mostly) STEM topics to individuals of varying knowledge levels: (1) child,

[1] https://www.wired.com/video/series/5-levels

#	Topic	#	Topic
1	Music harmony	14	Memory
2	Blockchain	15	Zero-knowledge proofs
3	Virtual reality	16	Black holes
4	Connectome	17	Quantum computing
5	Black holes	18	Quantum sensing
6	Lasers	19	Fractals
7	Sleep science	20	Internet
8	Dimensions	21	Moravecs Paradox
9	Gravity	22	Infinity
10	Computer hacking	23	Algorithms
11	Nanotechnology	24	Nuclear fusion
12	Origami	25	Time
13	Machine learning	26	Chess

Table 2: Topics in ReWIRED. 14-26 (yellow) are transcripts that were not part of the original WIRED dataset (Wachsmuth and Alshomary, 2022). The topic "black holes" is explained in two different videos, resulting in the duplicate (5, 16). Chess (26) applies distinctive knowledge levels (novice, intermediate, FIDE master, Grandmaster, and AI expert), as educational background doesn't imply a player's capability.

(2) teenager, (3) undergraduate college student, (4) graduate student, (5) colleague (another expert).

However, after the publication of the WIRED corpus, more clips came up in the video series. In our extension, ReWIRED, we incorporated all of them, doubling the original number data instances. In total, our dataset contains 130 transcripts from 26 topics across the five knowledge levels. Table 2 gives an overview of the topics covered.

3.2 Annotation Scheme: Teaching acts

We extend the dialogues by new span-level annotations of nine teaching acts, a dimension initially proposed by Feldhus et al. (2024). Table 1 lists the definitions of the teaching acts. Leveraging finer semantic granularity to teaching models in comparison to DAMSL (Core and Allen, 1997) and ISO 24617-2 (Bunt et al., 2012), the annotation framework we follow is similar to the CMA schema (Del-Bosque-Trevino et al., 2021), with further task-specific refinements.

We recruit real-world educational experts to incorporate domain-expert annotations in order to improve annotation quality and validity, particularly in modelling speaker interaction under instructional settings. Our four annotators are graduates with a Master of Education or similar, and with in-classroom teaching experience. They were paid at least the minimum wage in conformance with the standards of our host institutions' regions.

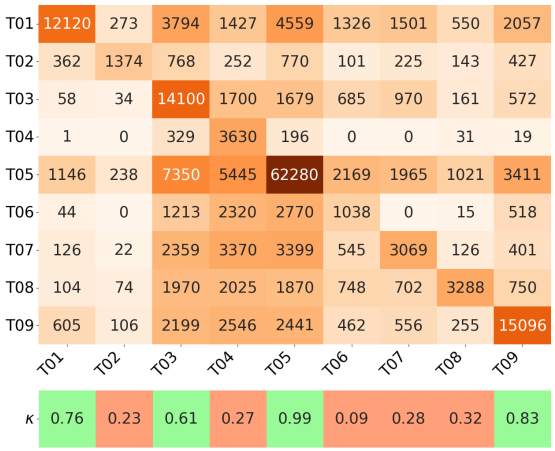
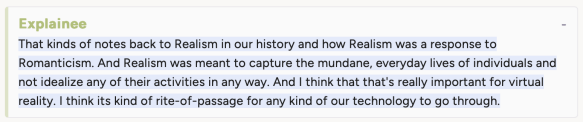
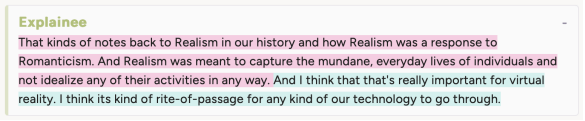


Figure 2: ReWIRED inter-annotator agreements for act on token level. For better visibility, we scale-adjust the colors by $\texttt{np.log1p}(\ldots)^3$. Each cell shows the number of tokens for which annotators (dis)agreed on a label in a pairwise comparison. The bottom row with green and red highlights show the Fleiss' κ per teaching act.



T05: Knowledge statement



T06: Comparison (pink) and T07: Generalization (azure)

Figure 3: An example of a turn given labeled as different teaching acts by the two expert annotators.

3.3 Annotation Process: Span Labeling

The full ReWIRED dataset is split into two fractions, each annotated by two out of four recruited expert annotators. As a post-processing step, we interpreted the inherited token-level annotations as a non-expert annotator and then consolidated all three annotations to yield the gold labels. Using three sets of annotations allows us to reduce the possibility of bias, especially in cases where two expert annotators disagree. The span-labeling task was performed on LABEL STUDIO (Tkachenko et al., 2020-2024), yielding an inter-annotator agreement of Fleiss' $\kappa = 0.44$. Figure 2 plots the respective agreement of the nine teaching act labels. An example is provided in Figure 3 to demonstrate how expert annotators could possibly disagree with each other on labeling a dialogue turn. Figure 4 shows the resulting distribution of teaching acts in our ReWIRED dataset.

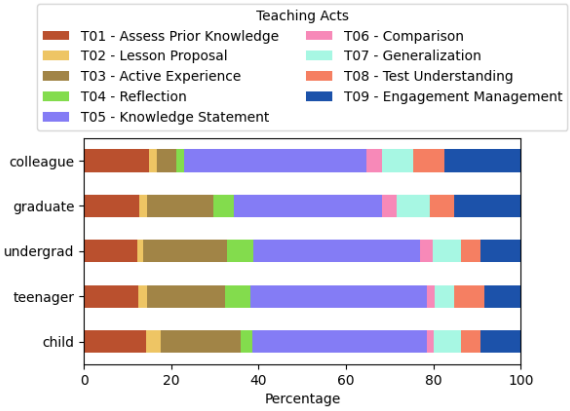


Figure 4: Distribution of teaching acts in ReWIRED across the five knowledge levels.

For quality evaluation, the annotators were also asked to evaluate instructional explanation categories with respect to the measurements of IXQUISITE (§5). Provided with respective descriptions, the annotators were asked to assess (a) presence and (b) contribution of each measurement on a 3-point Likert scale. This follow-up quality evaluation subsequent to span-labeling aimed to capture the views of educational experts across knowledge levels on the explanatory dialogues under the proposed framework. We discuss the outcome of this evaluation in Section 5.

4 Experiments: Sequence-labeling acts

To evaluate language models on detecting acts across act dimensions, we conduct experiments on span-labeling for ReWIRED, comparing the performance of a fine-tuned masked language model assigning token-level labels with that of LLMs. As a baseline, we follow Wachsmuth and Alshomary (2022) and evaluate `BERT` (Devlin et al., 2019) for token-level classification with 5-fold cross-validation, since the number of transcripts is not large enough to define partitions. We provide details on models in Appendix D.

We then frame the span-labeling task of the annotated labels (Table 1) in the ReWIRED dataset as a structured prediction task and analyze the capabilities of the following proprietary LLMs: `GPT-4o` (OpenAI, 2023), `Gemini 1.5 Flash`, and `Gemini 1.5 Pro` (Reid et al., 2024). In addition, we fine-tune `GPT-4o-mini` with 5-fold cross-validation (same setup as `BERT`, but with DPO, learning rate multiplier = 1.8, epochs = 3). We compare the following prompting approaches:

- **JSON**-type structured object prediction (Tavanaei et al., 2024; Wu et al., 2024);
- **TANL**-style structured prediction of inline tags (Paolini et al., 2021);
- **GoLLIE**-style information extraction using annotation guidelines (Sainz et al., 2024)

with the details further provided in Appendix E.

4.1 Results and discussion

The results for span-level act prediction (Table 3) reveal that the task remains rather challenging for LLMs, when they are not fine-tuned on the task. However, task performance could be significantly altered under the influence of the three prompting methods. First of all, we find the prompt design eliciting structured prediction in form of JSON objects to cause major problems for post-processing. The problematic output can nevertheless be mitigated by providing more context via few-shot demonstrations eliciting in-context learning: When including three previous dialogue turns and their gold labels, the predictions become more consistently structured (1.06% invalid JSONs by `GPT-4o`) and could achieve a noticeably higher performance. These findings reflect challenges reported by concurrent related works applying LLMs to dialogue tasks (Zhao et al., 2023) and span-labeling tasks (Ziems et al., 2024; Wang et al., 2023), and the difficulties of applying them to teaching settings (Wang and Demszky, 2023; Macina et al., 2023).

The outcomes of our experiments further highlight the impact of requested output format, as changing the structured prediction setup to TANL (bracketed tagging) or GoLLIE (coding guidelines) could almost double the performance of `GPT-4o` and `Gemini 1.5` across all nine teaching acts. Although the more straightforward TANL approach already yields consistent improvements leaving only T03 (*Active Experience*) and T08 (*Test Understanding*) slightly behind, the GoLLIE prompting method sees >80% micro-F_1 for the majority label (T05: *Knowledge Statement*). While `Gemini 1.5` performs best with the TANL approach, it could not use the GoLLIE paradigm well. `GPT-4o`'s generations are on a similar level between the two best prompting paradigms.

`BERT`, on the other hand, easily outperformed most LLMs used for inference across almost every single act. The stark difference can be attributed to the importance of fine-tuning and the constraint to predict one of the nine acts.

In our final set of experiments involving the fine-tuning of `GPT-4o-mini` with 5-fold cross-

Teaching acts	T01	T02	T03	T04	T05	T06	T07	T08	T09	Macro-F_1	Span Al.
BERT FT	**80.68** %	**72.15** %	**87.93** %	**83.07** %	**90.18** %	**81.57** %	**83.75** %	**82.53** %	**80.31** %	**84.17** %	–
GPT-4o JSON	35.69 %	49.38 %	39.80 %	34.60 %	66.36 %	38.76 %	39.34 %	29.19 %	42.72 %	41.76 %	36.75 %
GPT-4o TANL	66.69 %	70.39 %	63.61 %	80.22 %	84.91 %	75.10 %	75.29 %	61.96 %	70.26 %	72.05 %	68.21 %
GPT-4o GoLLIE	71.39 %	67.26 %	72.83 %	78.99 %	82.70 %	79.11 %	78.05 %	71.66 %	67.07 %	74.34 %	73.54 %
Gemini 1.5 F TANL	53.39 %	71.65 %	77.76 %	85.86 %	86.13 %	81.88 %	83.73 %	63.04 %	74.83 %	75.36 %	74.09 %
Gemini 1.5 F GoLLIE	46.17 %	45.95 %	59.33 %	69.39 %	72.82 %	64.41 %	65.47 %	47.84 %	49.89 %	57.92 %	58.80 %
Gemini 1.5 P TANL	67.11 %	74.00 %	79.97 %	79.45 %	87.18 %	81.35 %	82.03 %	53.70 %	77.51 %	75.71 %	69.81 %
Gemini 1.5 P GoLLIE	46.25 %	30.56 %	53.60 %	63.00 %	70.56 %	47.44 %	49.23 %	24.88 %	48.60 %	48.23 %	49.53 %
GPT-4o-mini FT TANL	93.64 %	97.98 %	95.23 %	**99.30** %	98.90 %	**99.03** %	98.64 %	97.00 %	97.28 %	97.44 %	94.63 %
GPT-4o-mini FT GoLLIE	**98.54** %	**98.57** %	**99.11** %	98.87 %	**99.56** %	98.14 %	**100.0** %	**99.67** %	**98.91** %	**99.04** %	**95.49** %

Table 3: Language models evaluated on the tasks of sequence-labeling teaching acts within dialogue turns from our ReWIRED dataset. Percentages under each of the acts show micro-F_1 scores in a 3-shot or fine-tuning (FT) setting. Span Alignment (last column) refers to how well the spans extracted by LLMs align with human-annotated spans.

validation, we can report that the gap between LLMs and BERT is non-existent, as fine-tuning can lead to further advances making it consistently annotate the teaching acts with up to 99.04% Macro-F_1 (GoLLIE) and 95.49% of spans correctly matched with human ground truth. For span-labeling tasks, we recommend practitioners to employ fine-tuning instead of few-shot prompting.

5 The IXQuisite test suite

Considering the interactive nature of dialogues, it is often challenging for human-free evaluation paradigms to cover criteria such as participation and engagement (Adiwardana et al., 2020) or capture conversational flow as perceived by human speakers (Deriu et al., 2021). In an attempt to map the linguistic features in dialogue form to the efficiency of explanations, Feldhus et al. (2024) introduced a didactic research-based test suite, with the name IXQUISITE. This test suite includes seven[2] teaching-act-related metrics that assess the *functional* content of explanations and seven additional metrics that evaluate the *form* of explanations. The metrics and their respective descriptions and references to their origins in the literature are presented in Table 4 and Table 5.

5.1 Human validation

First, to validate the alignment between our proposed metrics, hypotheses about their scores among the dialogues, and expert perception, we instructed our annotators to assess their presence and contribution in each dialogue they annotated. After completing each dialogue annotation (§3.2), they were asked to assess the presence of these metrics in the dialogue and how much they contribute to the explanation being suitable for the level of knowledge of the explainee. The annotators were provided with the descriptions in Tables 4 and 5 to ensure consistent evaluations across the dataset. The annotators had to assess each metric's presence and contribution, choosing between *non present*, *partially present* or *fully present*. The results of the annotators' assessment for the metrics' presence in the dataset are summarized in Figure 5a.

Our analysis indicates that the presence of the metrics, particularly function-oriented ones, strongly correlates with the five evaluation levels. This finding underscores the alignment of these metrics with the hierarchical framework of dialogue quality assessment. Interestingly, metrics such as *adaptation*, *readability*, and *coherence* demonstrated a uniform distribution of importance across the five levels, suggesting that these aspects are critical regardless of the specific level of quality of the dialogue. This observation may imply that these metrics are fundamental for explanatory dialogue evaluation, maintaining their relevance even as other aspects vary across levels.

5.2 Static evaluation on the dataset

Next, we directly applied the IXQUISITE metrics to our dataset's expert-annotated version. We use the automated, or *static*, metrics defined in Feldhus et al. (2024). Since our dataset lacks specific dialogue and explanation annotations, we excluded the *remedial explanations* metric from this evaluation and relied solely on T08 for the *check for understanding* category. For the *function-related* metrics, the number of tokens within each class represented in the metric is divided by the total number of tokens in the dialogue. Normalization is also applied to *minimal explanations*, *lexical complexity*, *syn-*

[2]In the original work, explanation-act annotations were also considered; since we do not have them in this work, we omit the metric of *remedial explanations* of our experiments.

IXQUISITE: Function metrics				
Abbr.	**Category**	**Description**	**Origin**	**Static metric**
PK	Check for prior knowledge	The teacher inquires the student about prior knowledge, background, or what their interests might be	Kulgemeyer and Schecker (2009), Leinhardt and Steele (2005)	T01
MI	Mindfulness of common misconceptions	The teacher addresses common misconceptions	Wittwer et al. (2010), Andrews et al. (2011)	T04
RE	Rule-example structure	The teacher states the abstract form of the concept being taught. Then, the teacher gives some examples to assist in understanding	Tomlinson and Hunt (1971)	T05 → T03
ER	Example-rule structure	For procedural knowledge, the teacher first provides examples and then derives the general rule from them	Champagne et al. (1982)	T03 → T05
EA	Example/Analogy connection	The teacher explains how parts of the analogy/example relate to the concept being explored	Ogborn et al. (1996), Valle and Callanan (2006)	T06
UN	Check for understanding	The teacher tests the understanding of the student	Webb et al. (1995)	T08

Table 4: Explanation and teaching acts-related measures in IXQUISITE for instructional explanation quality based on occurrences of classes from our annotation schema.

IXQUISITE: Form metrics				
Abbr.	**Category**	**Description**	**Origin**	**Static metric**
ME	Minimal explanations	Low cognitive load, e.g. avoid redundancies (verbosity) such as introducing named entities	Black et al. (1986)	Frequency of named entities
LC	Lexical complexity	The level of difficulty associated with any given word form by a particular individual or group	Kim et al. (2016)	Frequency of difficult words
SD	Synonym density	Children are proven better aligned with consistent terminology; experts allow more synonyms	Wittwer and Ihme (2014)	Frequency of synonyms for the n terms most connected to the topic
TM	Correlation to teaching model	Correlation of teaching act order to prescribed teaching models	Oser and Baeriswyl (2002), Krabbe et al. (2015)	Edit distance between T01-T08 (asc.) and actual occurrences
AD	Adaptation	The teacher incorporates prior knowledge, misconceptions and interests and uses analogies	Wittwer et al. (2010)	Inverse frequency of synonyms in the text
RL	Readability level	Indicator of how difficult a passage is to understand	Crossley et al. (2017)	Flesch-Kincaid Grade level
CO	Coherence	How sentences relate to each other to create a logical and meaningful flow for the reader or listener	Lehman and Schraw (2002), Duffy et al. (1986)	Frequency of conjunctions and linking language

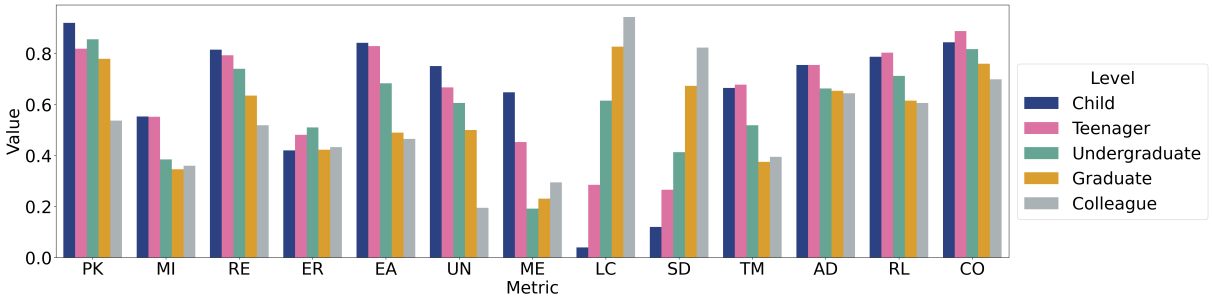
Table 5: Categories for instructional explanation quality and associated numerical measures in IXQUISITE.

onym density, and *coherence*. The results of this analysis are presented in Figure 5b. Our findings indicate that form-related metrics demonstrate a stronger correlation with the five predefined levels of knowledge, though the relationship is not perfectly linear. On the other hand, functional metrics related to teaching acts in our dataset only partially correlate with the five levels (see *PK*).

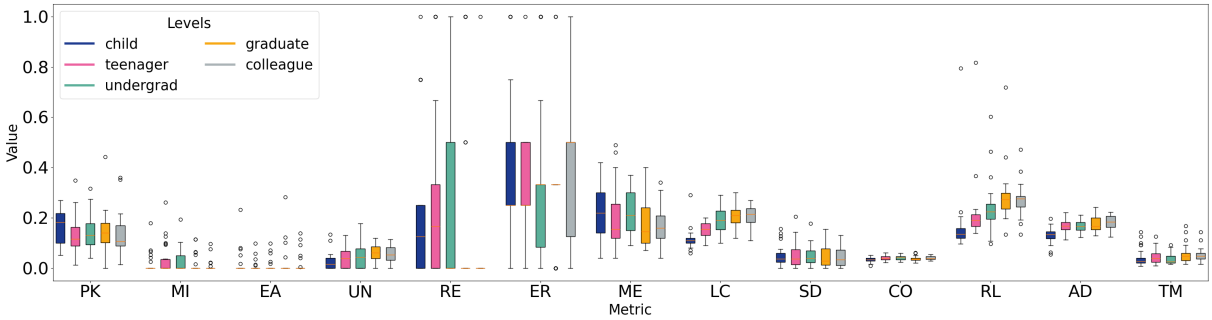
5.3 Prompt-based evaluation

Building on and inspired by the work of Rooein et al. (2024) on assessing readability across varying knowledge levels using *static* and *prompt-based* metrics, we extend this approach by formulating the metrics in the IXQUISITE suite as evaluative questions posed to a language model (specifically `GPT-4o`). Instead of designing closed-ended questions, we prompt `GPT-4o` to evaluate each metric on a scale from 0 to 10. For instance, rather than asking "Does the explainer inquire about prior knowledge?", we reframe the question as "On a scale from 0 to 10, how well does the explainer inquire about prior knowledge?". This approach facilitates the collection of more fine-grained results comparable to those provided by our human annotators.

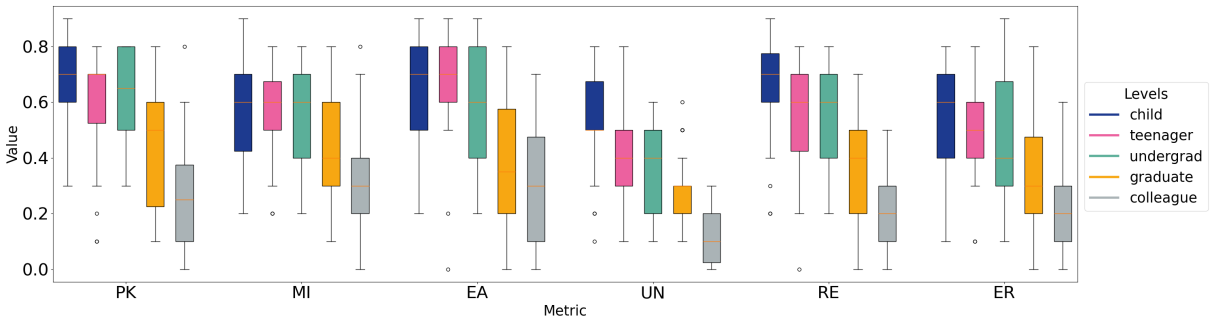
We present the outcomes of the prompt-based evaluation for function-related metrics (Table 4) in Figure 5c. The results for the form-related metrics (Table 5) can be found in Appendix F. We observe that function-related metrics exhibit the strongest correlation with the five levels of knowledge. Notably, the results obtained from prompt-based functional metrics align closely with human evaluations of the presence of each metric within individual dialogues, as there is greater variance across knowledge levels - typically showing a higher score range for lower knowledge levels than for the higher ones. This suggests that, while automated evaluation suffices for form-based metrics, capturing variation in functional metrics across dialogue levels requires prompt-based LLM evaluation.



(a) Annotators assessment on presence of each metric in IXQUISITE for in each level.



(b) IXQUISITE metrics: Static evaluation of our dataset.



(c) IXQUISITE function-related metrics: prompt-based evaluation of the five levels in the dataset.

Figure 5: IXQUISITE results.

6 Conclusion

In this paper, we have presented a dataset of explanatory instructional dialogues in one-to-one tutorial sessions, dubbed ReWIRED. In particular, we have extended the WIRED dataset (Wachsmuth and Alshomary, 2022), doubling the number of dialogues and adding span-level annotations of teaching acts reflecting practices according to teaching models in didactics literature. Our dataset has been annotated by teaching experts, with consolidated labels comparable to those of Feldhus et al. (2024).

With the annotated dataset, we have probed into the span-labeling task to classify teaching acts, conducting experiments on several language models of different sizes. The results disclosed that LLMs, including `GPT-4o` and `Gemini`, fall behind controlled setups with fine-tuning on a much smaller `BERT` or a `GPT-4o-mini` in reliably detecting teaching acts. Our findings inform future steps in operationalizing pedagogical theory for tutorial dialogues in NLP. They indicate that the IXQUISITE suite of metrics for assessing quality events in instructional explanations effectively captures the varying knowledge levels of the explainees. These metrics foster future work on automatically generating individualized and domain-specific explanations, contributing to the field of XAI and enhancing user experience.

Limitations

We acknowledge that, despite our annotators' high expertise in the field of education, some teaching acts seem not as easily distinguishable as the other act dimensions, resulting in a relatively low inter-annotator agreement. However, the single aggregation-based Fleiss' κ score might be too superficial to capture the complexity behind. Ultimately, the annotation variations also convey the subjectivity of teaching-related explanations, fol-

lowing the idea that human label variation should be encouraged (Plank, 2022).

Further limitations include that a portion of the test suite relies on human annotation, which may introduce inconsistencies. Replicating or extending the test suite might be difficult without a reliable teaching act prediction model. Also, the dataset we present is extracted from videos—audio and visual elements not present in the transcription. The efficacy of our approach may vary depending on the complexity and diversity of the multimodal inputs, if present. Last but not least, the generalizability of our findings may be constrained by the narrow domain of dialogues examined, limiting extrapolation to broader conversational contexts.

Ethical statement

We do not see immediate ethical concerns regarding research and development. The data included in the corpus are readily available from WIRED Web resources. Following the ACM Code of Ethics (1.2, 1.6), all participants consented to be recorded as far as perceivable from the WIRED web resources, which are free to use for research purposes. The two annotators in our study were recruited over online platforms (LinkedIn, university forum). The annotation of each dialogue took an annotator an average of 10 minutes; depending on their workload, the annotation duration was between 12 and 20 hours. In our view, the provided prediction models target dimensions of dialogue turns that are not prone to misuse for ethically doubtful applications.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Elaine Allensworth, Macarena Correa, and Steve Ponisciak. 2008. From high school to the future: Act preparation–too much, too late. why act scores are low in chicago and what it means for schools. *Consortium on Chicago School Research*.

Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. Modeling the quality of dialogical explanations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11523–11536, Torino, Italia. ELRA and ICCL.

Tessa M Andrews, Michael J Leonard, Clinton A Colgrove, and Steven T Kalinowski. 2011. Active learning not associated with student learning in a random sample of college biology courses. *CBE—Life Sciences Education*, 10(4):394–405.

John B. Black, John M. Carroll, and Stuart M. McGuigan. 1986. What kind of minimal instruction manual is the most effective. *SIGCHI Bull.*, 18(4):159–162.

Melissa Boston. 2012. Assessing instructional quality in mathematics. *The Elementary School Journal*, 113(1):76–104.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. Association for Computational Linguistics.

Audrey B Champagne, Leopold E Klopfer, and Richard F Gunstone. 1982. Cognitive research and the design of science instruction. *Educational Psychologist*, 17(1):31–53.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Jorge Del-Bosque-Trevino, Julian Hough, and Matthew Purver. 2021. Communicative grounding of analogical explanations in dialogue: A corpus study of conversational management acts and statistical sequence models for tutoring through analogy. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 23–31, Gothenburg, Sweden. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

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.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.*, 54(1):755–810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Gerald G. Duffy, Laura R. Roehler, Michael S. Meloth, and Linda G. Vavrus. 1986. Conceptualizing instructional explanation. *Teaching and Teacher Education*, 2(3):197–214.

Nils Feldhus, Aliki Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, page 225–230, New York, NY, USA. Association for Computing Machinery.

Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine L. Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, Roni Rabin, Jasmin Rubinovitz, Amit Pitaru, Mac McAllister, Julia Wilkowski, David Choi, Roee Engelberg, Lidan Hackmon, Adva Levin, Rachel Griffin, Michael Sears, Filip Bar, Mia Mesar, Mana Jabbour, Arslan Chaudhry, James Cohan, Sridhar Thiagarajan, Nir Levine, Ben Brown, Dilan Görür, Svetlana Grant, Rachel Hashimshoni, Laura Weidinger, Jieru Hu, Dawn Chen, Kuba Dolecki, Canfer Akbulut, Maxwell L. Bileschi, Laura Culp, Wen-Xin Dong, Nahema Marchal, Kelsie Van Deman, Hema Bajaj Misra, Michael Duah, Moran Ambar, Avi Caciularu, Sandra Lefdal, Christopher Summerfield, James An, Pierre-Alexandre Kamienny, Abhinit Mohdi, Theofilos Strinopoulos, Annie Hale, Wayne Anderson, Luis C. Cobo, Niv Efron, Muktha Ananda, Shakir Mohamed, Maureen Heymans, Zoubin Ghahramani, Yossi Matias, Ben Gomes, and Lila Ibrahim. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *CoRR*, abs/2407.12687.

Maria Kasinidou, Styliani Kleanthous, and Jahna Otterbacher. 2024. "artificial intelligence is a very broad term": How educators perceive artificial intelligence? In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, page 315–323, New York, NY, USA. Association for Computing Machinery.

Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical simplification of scientific terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1071, Austin, Texas. Association for Computational Linguistics.

Heiko Krabbe, Simon Zander, and Hans Ernst Fischer. 2015. *Lernprozessorientierte Gestaltung von Physikunterricht - Materialien zur Lehrerfortbildung*. Waxmann.

Christoph Kulgemeyer. 2018. Towards a framework for effective instructional explanations in science teaching. *Studies in Science Education*, 54(2):109–139.

Christoph Kulgemeyer and Horst Schecker. 2009. Kommunikationskompetenz in der physik: Zur entwicklung eines domänenspezifischen kompetenzbegriffs. *Zeitschrift für Didaktik der Naturwissenschaften*, 15:131–153.

Stephen Lehman and Gregory Schraw. 2002. Effects of coherence and relevance on shallow and deep text processing. *Journal of Educational Psychology*, 94(4):738–750.

Gaea Leinhardt and Michael D. Steele. 2005. Seeing the complexity of standing to the side: Instructional dialogues. *Cognition and Instruction*, 23(1):87–163.

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

Jeanne McClure, Machi Shimmei, Noboru Matsuda, and Shiyan Jiang. 2024. Leveraging prompts in llms to overcome imbalances in complex educational text data. *arXiv*, abs/2407.01551.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 225–235. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Jon Ogborn, Gunther Kress, Isabel Martins, and Kieran McGillicuddy. 1996. *Explaining science in the classroom*. McGraw-Hill Education (UK).

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Fritz Oser and Franz Baeriswyl. 2002. *AERA's Handbook of Research on Teaching, 4th Edition*, pages 1031–1065. Washington: American Educational Research Association (AERA).

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Hendrik Schuff, Heike Adel, Peng Qi, and Ngoc Thang Vu. 2023. Challenges in explanation quality evaluation. *arXiv*, abs/2210.07126v2.

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.

Amir Tavanaei, Kee Kiat Koo, Hayreddin Ceker, Shaobai Jiang, Qi Li, Julien Han, and Karim Bouyarmane. 2024. Structured object language modeling (SO-LM): Native structured objects generation conforming to complex schemas with self-supervised denoising. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 821–828, Miami, Florida, US. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2024. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Peter D Tomlinson and David E Hunt. 1971. Differential effects of rule-example order as a function of learner conceptual level. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 3(3):237.

Araceli Valle and Maureen A Callanan. 2006. Similarity comparisons and relational analogies in parent-child conversations about science topics. *Merrill-Palmer Quarterly (1982-)*, pages 96–124.

Henning Wachsmuth and Milad Alshomary. 2022. "mama always had a way of explaining things so I could understand": A dialogue corpus for learning to construct explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 344–354, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Rose E. Wang, Ana T. Ribeiro, Carly Robinson, Susanna Loeb, and Dora Demszky. 2024. Tutor CoPilot: A

human-AI approach for scaling real-time expertise. *arXiv*, abs/2410.03017.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *arXiv*, abs/2304.10428.

Noreen M Webb, Jonathan D Troper, and Randy Fall. 1995. Constructive activity and learning in collaborative small groups. *Journal of educational psychology*, 87(3):406.

Jörg Wittwer, Matthias Nückles, Nina Landmann, and Alexander Renkl. 2010. Can tutors be supported in giving effective explanations? *Journal of Educational Psychology*, 102(1):74.

Jörg Wittwer and Natalie Ihme. 2014. Reading skill moderates the impact of semantic similarity and causal specificity on the coherence of explanations. *Discourse Processes*, 51(1-2):143–166.

Haolun Wu, Ye Yuan, Liana Mikaelyan, Alexander Meulemans, Xue Liu, James Hensman, and Bhaskar Mitra. 2024. Learning to extract structured entities using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6817–6834, Miami, Florida, USA. Association for Computational Linguistics.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4375–4389. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT equipped with emotional dialogue capabilities? *arXiv*, abs/2304.09582.

Zining Zhu and Frank Rudzicz. 2023. Measuring information in text explanations. *CoRR*, abs/2310.04557.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational linguistics*, 50(1).

Appendix

A Annotation instructions

To annotators, we provided examples from Appendix B as well as further delineations of the acts with examples and descriptions of how to differentiate between them. We also provided a screencast with instructions on how to use LABEL STUDIO and walk-through examples for each act. This will be published with the camera-ready version. The introductory text shown to all annotators before watching the recording and accessing LABEL STUDIO is the following (unformatted version):

> Your objective is annotating linguistic information about the multi-layered objectives each person performs when communicating. The dataset is comprised of transcribed conversations in which an expert in a field explains some concept to multiple people at varying levels of education: child, teenager, undergraduate, graduate and expert.
> Your task as an annotator will be, given a transcript of one of these conversations, to use a highlighting tool to mark which "acts" are present in different parts of the text. These acts highlight some unspoken objectives present in the text. For example, the text "Do you understand that?" could be said to have both an objective of asking a yes/no question and checking for understanding.
> Some of these will be straightforward to label and say "that is clearly the intention behind that sentence", while some will be a bit more complicated. We often have many intentions behind what we say, and we account for that by letting you tag any segment of text with as many labels as you see fit, even none at all.
> Your annotation task is about labeling the aforementioned objectives from the perspective of Teaching Acts, which focus on conversation mechanics in terms of lesson planning and didactics.

B Examples for acts

Figure 6 shows examples from ReWIRED for each of the acts as provided to the annotators.

C Label distributions

Figure 7 shows the number of distinct acts per dialogue turn as per annotated.

D Models

Table 6 lists how the models in §4 were employed. We used the following GPUs: A100, RTXA6000, RTX3080. For the BERT fine-tuning, we reinitialized the BERT model for token classification at the start of every fold ($k = 5$) and used a batch size of 4, an AdamW optimizer with a learning rate of $5 * 10^{-6}$, epsilon of $1 * 10^{-8}$, and warmup.

E Prompt design

Figure 8 and Figure 9 depict the prompts used with LLMs such as `GPT-4o` to produce the predictions whose evaluation is shown in Table 3. For few-shot demonstrations, we first presented the three preceding turns of the same dialogue (or from the end of

Model name	#Params	URL	Training times	Inference times
BERT	110M	https://huggingface.co/bert-base-uncased	13 hours	<1 hour
GPT-4o-mini (fine-tuned)	?	https://platform.openai.com/docs/guides/fine-tuning	6 hours	6 hours
GPT-4o	?	https://platform.openai.com/docs/api-reference/chat	n.a.	9 hours
Gemini 1.5	?	https://ai.google.dev/gemini-api/docs	n.a.	11 hours

Table 6: Language models with parameter counts, training times, inference times, and API costs.

last dialogue if the turn in question is at the start of a dialogue) and their corresponding gold spans (in the format required by the respective prompting paradigm) just as we elicit it from the model in the zero-shot setup. Figure 10 and Figure 11 show the results from GoLLIE and TANL prompts for Gemini 1.5 Pro and GPT-4o, respectively.

F IXQuisite: additional information

F.1 Annotator's assessment of contribution of metrics in each level

Besides validating the presence of each IXQUISITE metric in every dialogue, annotators were additionally asked to assess their importance/contribution, especially in regards to the level of knowledge of the explainee. Figure 12 shows the annotator's assessment of the importance/contribution of each metric at each level.

F.2 Form metrics: prompt-based evaluation

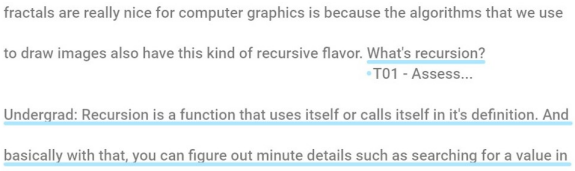
sFigure 13 presents the results of the prompt-based evaluation of the form metrics in the dataset. The results do not exhibit a clear correlation with the five levels, predominantly falling within the range of 0.8 to 0.9. This may be attributed to the formulation of the prompts.0.9. This might be related to the way the prompts were formulated.

F.3 Prompt-based metric questions

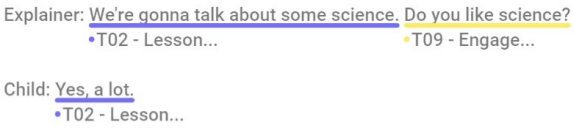
Table 7 shows the metrics formulated as questions for prompt-based evaluation of the explanatory dialogues in the ReWIRED dataset according to the IXQUISITE test suite.

Abbr.	On a scale from 0 to 10...
PK	... how well does the explainer inquire about prior knowledge?
MI	... how well does the explainer deal with common misconceptions?
RE	... how well does the explainer state the abstract form of a statement and then some example to assist understanding?
ER	... how well does the explainer provide examples prior to deriving a rule?
EA	... how well does the explainer explain ... how parts of the analogy/example relate to the concept being explored?
UN	... how well does the explainer check the understanding of the student?
ME	... how appropriate is the cognitive load for the explainee's level?
LC	... how appropriate is the lexical complexity for the explainee's level?
SD	... how appropriate is the amount of synonyms and technical language used for the explainee's level?
AD	... how well-adapted is the content of the dialogue to the explainee?
RG	... how appropriate is the readability level for the explainee's level?
CO	... how appropriate is the number of conjuction and subordination for the explainee's level?
TM	... how coherent is the text for the explainee's level?"

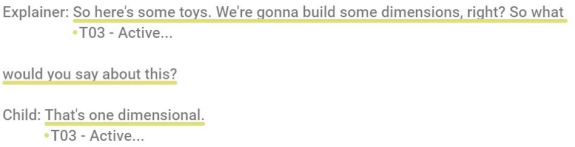
Table 7: IXQUISITE metrics formulated as questions for prompt-based dialogue evaluation.



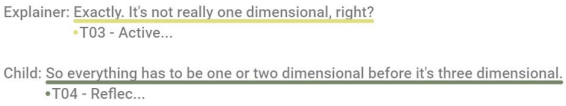
(a) T01: Assess Prior Knowledge



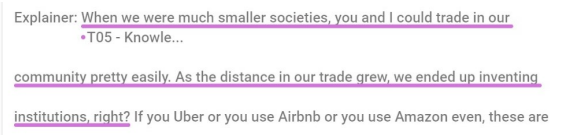
(b) T02: Lesson Proposal



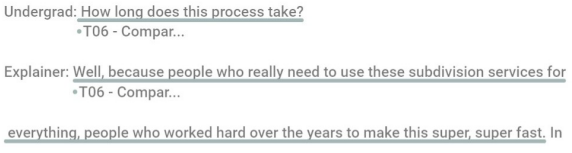
(c) T03: Active Experience



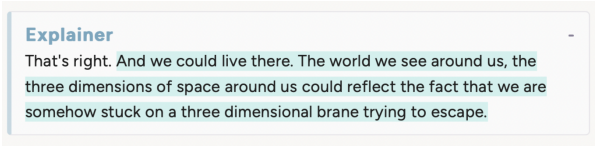
(d) T04: Reflection



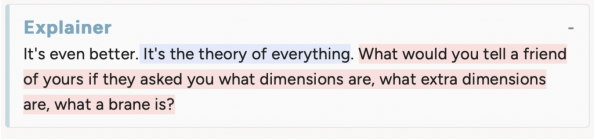
(e) T05: Knowledge Statement



(f) T06: Comparison



(g) T07: Generalization



(h) T05: Knowledge Statement (blue) and T08: Test Understanding (vermilion)

(i) T09: Engagement Management

Figure 6: Examples for teaching acts T01-T09.

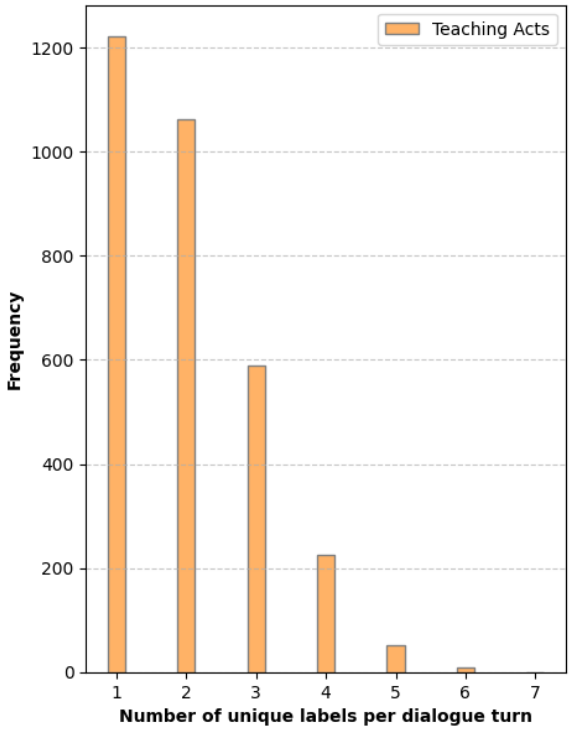


Figure 7: Number of unique teaching acts per turn in ReWIRED. The bar chart reveals that more than half of all dialogue turns in ReWIRED contain more than one distinct teaching act.

```python
# Example label mapping (dialogue acts)
ReWIRED_ta_str_2_int = {
    'T01 - Assess Prior Knowledge': 1,
    'T02 - Lesson Proposal': 2,
    'T03 - Active Experience': 3,
    'T04 - Reflection': 4,
    'T05 - Knowledge Statement': 5,
    'T06 - Comparison': 6,
    'T07 - Generalization': 7,
    'T08 - Test Understanding': 8,
    'T09 - Engagement Management': 9,
    'T10 - Other Act': 0
}
label_schema = ("The label schema consists of the following 10 classes:\n* " + "\n*
↪  ".join(list(ReWIRED_ta_str_2_int.keys())) + "\n")
```

Figure 8: Label schema.

```python
system_prompt = (f"You are an expert annotator. ")
read_instruction = (f"Here is one turn from a dialogue between an explainer and a {student_role}
↪  on the topic of {topic}:\n{turn_text}\n")

task_instruction_JSON = ("Please extract the spans from the turn and assign a label to each of
↪  the spans. It is possible that the whole turn is just one span, because the act applies to
↪  its entirety. Please present your predictions in a JSON format like this:
↪  {\n\t{\n\t\t'Span': '...', \n\t\t'Predicted label': '...' \n\t},\n}\n")
task_instruction_TANL = ("Please annotate the spans in the turn by marking them inline using the
↪  format [ span | label ]. It is possible that the whole turn is just one span if the act
↪  applies to its entirety.")
task_instruction_GoLLIE = ("Task: Annotate the following text with {TASK_NAME[task]}
↪  labels.\n\n'docstring += 'Guidelines:\n'docstring += '- Identify spans in the text that
↪  correspond to the following acts.\n'docstring += '- The act classes are defined below.")

entire_input = system_prompt + read_instruction + label_schema + task_instruction
```

Figure 9: Simplified version of the Python code showing the span-labeling task prompt for ReWIRED.

```python
Text = "Explainer: \"So machine learning is a way that we teach computers to learn things about
↪  the world by looking at patterns and looking at examples of things. So can I show you an
↪  example of how a machine might learn something?\""

labels = [
    {'span': "So machine learning is a way that we teach computers to learn things about the
    ↪  world by looking at patterns and looking at examples of things.", 'label':
    ↪  'T05___Knowledge_Statement'},
    {'span': "So can I show you an example of how a machine might learn something?", 'label':
    ↪  'T02___Lesson_Proposal'},
]
```

Figure 10: Example for a result from a GoLLIE prompt with `Gemini 1.5 Pro`.

```
"Explainer: ""It's a lot of practice and analysis. [Really, an advanced chess player was not
↪  born an advanced chess player. They have probably hundreds, if not thousands of more games
↪  in their mind, in their past, in their history that they've analyzed, that they've studied.
↪  It's like any athlete, you know? | T07 - Generalization] [I put my weight on this foot, and
↪  so I wasn't able to hit the shot back that well. So the next time that that happens, I'm
↪  gonna be more prepared. | T06 - Comparison]"""
```

Figure 11: Example for a result from a TANL prompt with `GPT-4o`.

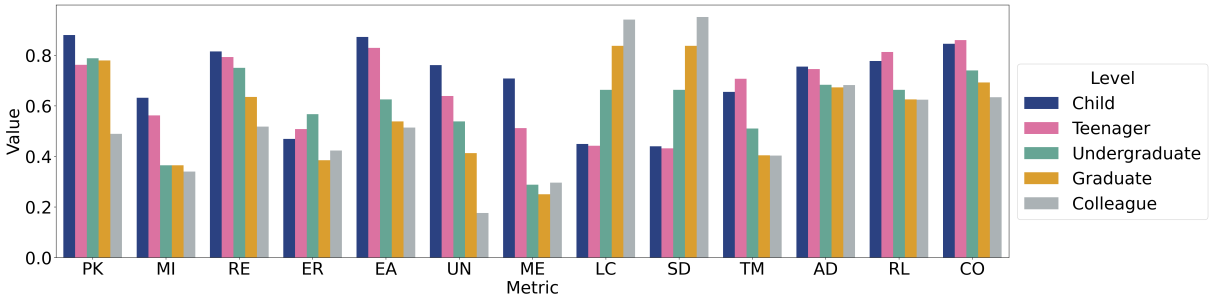


Figure 12: Annotators assessment on contribution of each metric present in IXQUISITE for each level.

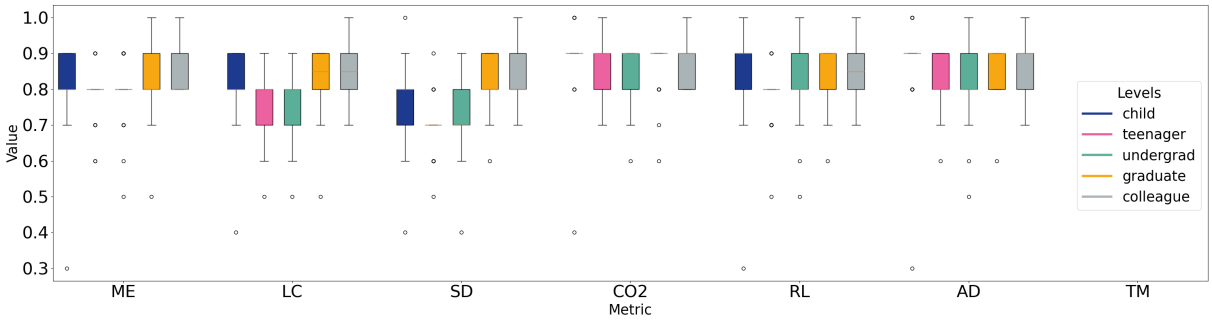


Figure 13: IXQUISITE form metrics: prompt-based evaluation of the five levels in the dataset.