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

Aliki Anagnostopoulou[*] **Nils Feldhus**[1,3,4] **Yi-Sheng Hsu**[*,7]
Milad Alshomary[5] **Henning Wachsmuth**[6] **Daniel Sonntag**[1,2]

[1]German Research Center for Artificial Intelligence (DFKI) [2]Applied Artificial Intelligence, Oldenburg University
[3]Technische Universität Berlin [4]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[5]Data Science Institute, Columbia University [6]Leibniz University Hannover, Institute of Artificial Intelligence
[7]Computer Science Institute, Ruhr West University of Applied Sciences

Corresponding authors: feldhus@tu-berlin.de h.wachsmuth@ai.uni-hannover.de

Abstract

Understanding the strategies that make expert-led explanations effective is a core challenge in didactics and a key goal for explainable AI. To study this computationally, we introduce ReWIRED, a large corpus of explanatory dialogues annotated by education experts with fine-grained, span-level teaching acts across five levels of explainee knowledge. We use this resource to assess the capabilities of modern language models, finding that while few-shot LLMs struggle to label these acts, fine-tuning is a highly effective methodology. Moving beyond structural annotation, we propose and validate a suite of didactic quality metrics. We demonstrate that a prompt-based evaluation using an LLM as a "judge" is required to capture how the functional quality of an explanation aligns with the learner's expertise – a nuance missed by simpler static metrics. Together, our dataset, modeling insights, and evaluation framework provide a comprehensive methodology to bridge pedagogical principles with computational discourse analysis.

1 Introduction

Effective teaching is a masterclass in communication, where an expert dynamically adapts their language and strategy to guide a learner toward understanding. This process unfolds as a complex, structured dialogue, yet the specific discourse mechanisms that make an explanation effective, especially when tailored to different audiences, are not well understood from a computational perspective. While insights from education and psychology define what constitutes good teaching (Miller, 2019; Kulgemeyer, 2018), we lack the fine-grained datasets and evaluation frameworks needed to model these principles in natural language.

This paper addresses that gap through a multi-faceted approach, as illustrated in Figure 1. First, we introduce ReWIRED, a new corpus resource that significantly extends the WIRED dataset (Wachsmuth and Alshomary, 2022). Our contribution lies in a new layer of **span-level annotations of teaching acts**, provided by education domain experts, across dialogues tailored to five distinct knowledge levels (from child to colleague). This provides an empirical foundation for studying pedagogical discourse structure (§3).

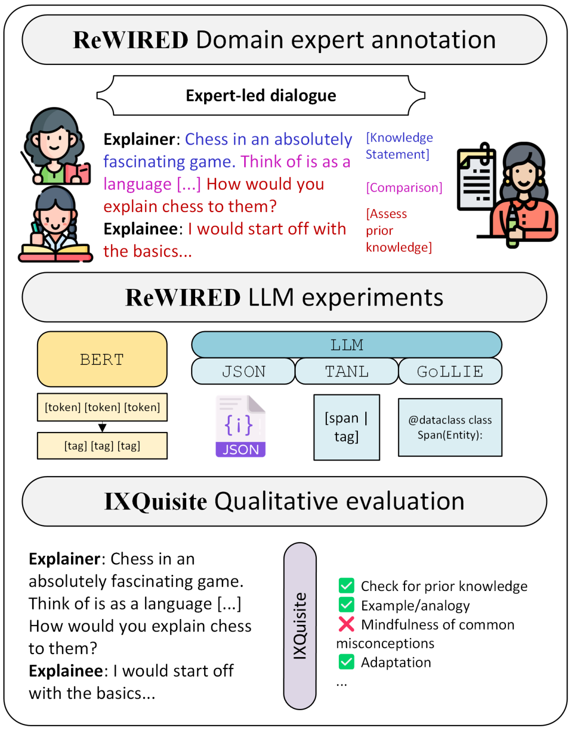

ReWIRED Domain expert annotation
Expert-led dialogue
Explainer: Chess in an absolutely fascinating game. Think of is as a language [...] How would you explain chess to them?
Explainee: I would start off with the basics...
[Knowledge Statement]
[Comparison]
[Assess prior knowledge]
ReWIRED LLM experiments
BERT
[token] [token] [token]
[tag] [tag] [tag]
LLM
JSON
TANL
GoLLIE
{i}
JSON
[span | tag]
@dataclass class Span(Entity):
IXQuisite Qualitative evaluation
Explainer: Chess in an absolutely fascinating game. Think of is as a language [...] How would you explain chess to them?
Explainee: I would start off with the basics...
IXQuisite
✓ Check for prior knowledge
✓ Example/analogy
✗ Mindfulness of common misconceptions
✓ Adaptation
...

Figure 1: Our workflow: We begin by having education experts create span-level annotations of teaching acts in explanatory dialogues. We then experiment with various LLMs to automate this annotation. Finally, we conduct a qualitative evaluation, using both human experts and LLMs, to assess the quality of the explanations based on didactic principles.

Second, we explore the feasibility of automating the detection of these acts. We evaluate a range of language models and prompting techniques, reveal-

[*] Work done while at DFKI.

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 166–181
November 9, 2025 ©2025 Association for Computational Linguistics

ing that while few-shot LLMs struggle with this nuanced task, models fine-tuned on our data—even smaller ones—can achieve near-perfect accuracy. This establishes a robust methodology for analyzing instructional dialogues at scale (§4).

Finally, we move from structural annotation to quality assessment. We employ and extend IXQUISITE, a suite of metrics grounded in didactics, to evaluate explanation quality. We validate these metrics with our expert annotators and demonstrate that a prompt-based evaluation using LLMs as "judges" is significantly more effective at capturing the functional quality of instructional discourse than traditional static methods. This provides a new paradigm for evaluating pedagogically-aware systems (§5).

Together, these contributions – a richly annotated corpus, a validated modeling approach, and a nuanced evaluation framework – provide a comprehensive methodology for bridging educational theory with computational discourse analysis, paving the way for AI systems that can generate more effective, human-like explanatory dialogues [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, we focus on its execution as a social, dialogical practice (Miller, 2019). In this view, the sequence of communicative acts, the choice of examples, and the adaptation to the learner are all crucial elements of the dialogue's discourse structure.

Modeling Pedagogical Strategies with Annotation Schemata. To analyze this structure computationally, we draw from established *teaching models* from education science (Oser and Baeriswyl, 2002; Krabbe et al., 2015). These models are not just abstract theories; they provide a blueprint for effective instructional sequences. For instance, a common pattern is to first assess prior knowledge, then introduce a concept, provide an example, and finally test for understanding. We operationalize these pedagogical principles as a set of nine span-level *teaching acts* (Table 1). This approach treats teaching strategies as a form of domain-specific discourse annotation, allowing us to model the underlying functional structure of the dialogue beyond surface-level linguistics.

Corpora for Educational Dialogue and Explanation Quality. Several corpora have paved the way for analyzing educational dialogues. Datasets like CIMA (Stasaski et al., 2020), TSCC-2 (Caines et al., 2022), and NCTE (Demszky and Hill, 2023) capture teacher-student interactions, but often focus on general dialogue moves rather than the specific pedagogical functions within an explanation. The work closest to ours is the WIRED corpus (Wachsmuth and Alshomary, 2022) and its analysis by Alshomary et al. (2024), which includes annotations for high-level explanation and dialogue moves. Our work significantly extends this by: (1) doubling the dataset size; (2) providing more granular, **span-level** annotations of teaching acts rather than turn-level classifications; and (3) using **domain experts** in education for annotation, increasing the validity of the labels. This finer granularity is crucial for understanding how different teaching strategies are woven together within a single conversational turn.

Recent work has also leveraged LLMs in education, for tasks like assessing student answers (Carpenter et al., 2024) or cognitive engagement (McClure et al., 2024), and in human-AI tutoring systems (Wang et al., 2024; Jurenka et al., 2024). Evaluating the quality of these interactions remains a challenge. While some metrics focus on general dialogue quality (Mehri and Eskénazi, 2020) or textual features (McNamara et al., 2014), they often miss the pedagogical dimension. Inspired by the approach of Rooein et al. (2024), who use both static and LLM-prompted metrics for readability, we adopt and expand a suite of quality metrics to specifically assess instructional explanations, connecting discourse phenomena to didactic principles. This addresses the challenge noted by Xu et al.

[1] The dataset, code, and test suite are available at https://github.com/nfelnlp/InstruX.

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).

(2024) that LLMs excel at simple evaluation but struggle with complex teaching practices without proper guidance.

3 The ReWIRED dataset

To study instructional strategies in explanatory dialogues, we introduce ReWIRED, a new corpus resource featuring a novel layer of expert-provided, span-level annotations. We build upon and significantly extend an existing dataset of instructional dialogues, enriching it with annotations grounded in pedagogical theory to facilitate fine-grained discourse analysis.

3.1 Source data: Explanation dialogues

Our starting point is the WIRED corpus (Wachsmuth and Alshomary, 2022), which contains transcripts from the *5-Levels* video series[2]. These videos provide a unique setting for discourse analysis, as they feature a domain expert explaining a complex STEM topic to five different explainees of progressively higher expertise: (1) a child, (2) a teenager, (3) an undergraduate, (4) a graduate student, and (5) a colleague (a fellow expert).

#	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.

We expanded this resource by transcribing and incorporating 13 additional topics released after the original corpus' publication, effectively doubling the dataset size. ReWIRED now comprises 130 dialogues across the 26 topics shown in Table 2. This expansion broadens the dataset's scope and enriches the variety of linguistic phenomena available for analysis.

3.2 Annotation of Teaching Acts

The primary contribution of our work is a new layer of annotation. We argue that to model how instruction is delivered, we need annotations that are more granular than turn-level labels. Pedagogical strategies are often embedded within a single utterance or can overlap. Therefore, we adopt a **span-labeling** approach to precisely identify segments corresponding to nine distinct *teaching acts*, as defined in Table 1. This annotation scheme allows us to capture the fine-grained, and often nested, discourse structure of instructional explanations.

Annotation Process and Quality. To ensure the validity of our annotations, we recruited four annotators, all of whom hold a Master of Education degree or equivalent and have practical in-classroom teaching experience. Annotators were onboarded through a detailed process that included a written guide with definitions and examples for each act (see Appendix E and A), and a screencast

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

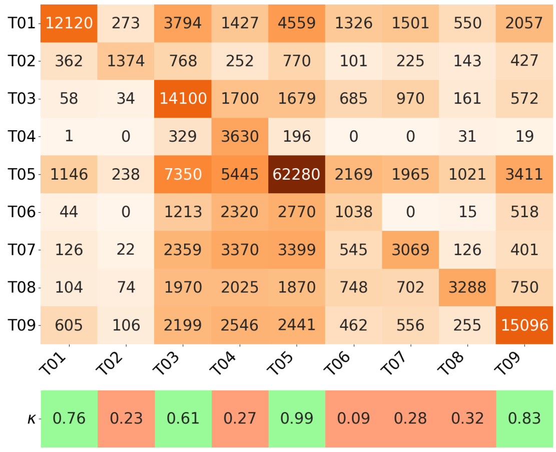


Figure 2: ReWIRED inter-annotator agreement for teaching acts on token level. For better visibility, we scale-adjust the colors by `np.log1p(...)`^{3}. Each cell shows the number of tokens for which annotators (dis)agreed on a label in a pairwise comparison. The bottom row with green and red highlights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

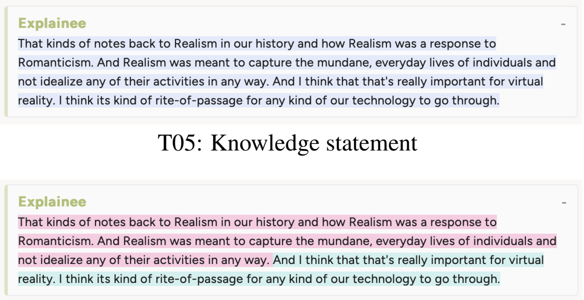
demonstrating the annotation tool (Label Studio (Tkachenko et al., 2020-2024)) and walking through ambiguous cases.

The full dataset was split, with each half annotated by two experts. The task proved to be challenging, reflecting the inherent subjectivity of interpreting pedagogical intent. This is visible in Figure 3, which shows how two experts can reasonably apply different labels to the same text. The resulting inter-annotator agreement is Fleiss' $\kappa = 0.44$. While this value indicates moderate agreement, it is not unexpected for a complex discourse annotation task and highlights that human label variation can itself be an informative signal about the ambiguity of the underlying phenomena (Plank, 2022). To create a reliable gold standard, we introduced the pre-existing non-expert annotations from Feldhus et al. (2024) as a third opinion and consolidated all three label sets to adjudicate disagreements.

The final distribution of teaching acts across the five knowledge levels is shown in Figure 4. This newly annotated corpus provides a unique resource for studying how discourse strategies in explanations are adapted to listeners with varying levels of prior knowledge.

4 Experiments: Sequence-labeling acts

Having established a richly annotated dataset, a critical next step is to assess the feasibility of automating the detection of teaching acts. Automating this process is a prerequisite for analyzing instructional discourse at scale or for developing real-time assistive technologies. We therefore conduct a series of experiments to evaluate how well modern language models can perform this complex, span-level sequence labeling task.


Explainee

That kinds of notes back to Realism in our history and how Realism was a response to Romanticism. And Realism was meant to capture the mundane, everyday lives of individuals and not idealize any of their activities in any way. And I think that that's really important for virtual reality. I think its kind of rite-of-passage for any kind of our technology to go through.

T05: Knowledge statement

Explainee

That kinds of notes back to Realism in our history and how Realism was a response to Romanticism. And Realism was meant to capture the mundane, everyday lives of individuals and not idealize any of their activities in any way. And I think that that's really important for virtual reality. I think its kind of rite-of-passage for any kind of our technology to go through.

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.

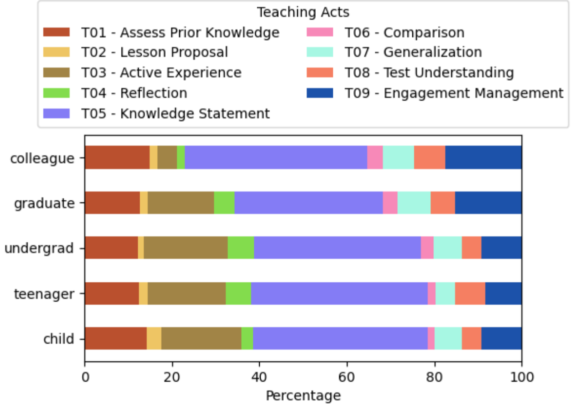


Figure 4: Distribution of teaching acts in ReWIRED across the five knowledge levels.

We frame the task as structured prediction on the ReWIRED dialogues. Our evaluation compares three distinct approaches: a fine-tuned baseline model, large language models (LLMs) in a few-shot setting, and a fine-tuned LLM.

Models and Setups. As a strong baseline, we fine-tune `BERT-base` (Devlin et al., 2019) for token-level classification using 5-fold cross-validation, following the setup of Wachsmuth and Alshomary (2022). We then evaluate large proprietary LLMs—`GPT-4o` (OpenAI, 2023) and two versions of `Gemini 1.5` (Reid et al., 2024)—using few-shot prompting. Finally, to directly compare the effect of fine-tuning on a modern architecture, we fine-tune `GPT-4o-mini` using the same 5-fold cross-validation setup. Further details on model implementation are in Appendix C.

Prompting for Structured Prediction. For the LLM experiments, we test three different prompt-

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.

ing paradigms designed to elicit structured, span-level output:

- **JSON**: Requesting a list of JSON objects, each containing a text span and its predicted label (Wu et al., 2024).
- **TANL**: An inline tagging format where predictions are structured as `[span | label]` directly in the text (Paolini et al., 2021).
- **GoLLIE**: Generating Python-like code where spans and labels are assigned to data structures, guided by a schema provided in the prompt (Sainz et al., 2024).

GPT-4o-mini is fine-tuned with 5-fold cross-validation (same setup as BERT, but with DPO, learning rate multiplier = 1.8, epochs = 3). Details and examples of these prompts are provided in Appendix D.

4.1 Results and discussion

Our experimental results, presented in Table 3, reveal several key insights into modeling domain-specific discourse acts.

Few-shot LLMs struggle with structured output and complex acts. Without fine-tuning, LLMs find the task challenging. The **JSON** format proved particularly unreliable, frequently producing malformed output that complicated post-processing and led to poor performance. While providing few-shot examples improved output consistency, the overall results remained low. Switching to more constrained output formats like **TANL** and **GoLLIE** yielded substantial improvements, nearly doubling the Macro-F_1 for GPT-4o. This highlights that for complex structured prediction, the choice of output format is critical. Even so, performance varied substantially across models and prompting styles, with TANL emerging as the best few-shot approach, but still lagging behind the exceptional performance of fine-tuning.

Fine-tuning is essential for high performance. The fine-tuned BERT baseline handily outperformed all few-shot LLM configurations across nearly every teaching act. This underscores the difficulty of the task and suggests that successfully capturing nuanced, domain-specific discourse phenomena requires task-specific adaptation.

This conclusion is further reinforced by our final experiment: the fine-tuned GPT-4o-mini achieves near-perfect scores, with a Macro-F_1 of up to 99.04% and a span alignment of 95.49%. Rather than suggesting the task is trivial, this result demonstrates that **fine-tuning is the most effective and reliable paradigm for this task**. It shows that even a smaller, more efficient LLM, when properly adapted with in-domain data, can master the complexities of annotating pedagogical discourse. For practitioners seeking to automate the analysis of such dialogues, we strongly recommend fine-tuning over few-shot prompting.

5 The IXQuisite test suite

While our experiments show that teaching acts can be reliably annotated with fine-tuning, the presence of individual acts does not guarantee a high-quality explanation. A good instructional dialogue must orchestrate these acts into a coherent and effective structure. Evaluating this holistic quality is challenging for standard automated metrics, which often fail to capture the nuances of conversational flow and engagement (Deriu et al., 2021).

To address this, we employ and extend IXQuisite, a test suite of quality metrics for instructional explanations grounded in didactic research (Feldhus et al., 2024). The metrics are di-

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. The right arrow between two teaching acts in static metrics refers to passages where two different acts directly follow one another in this exact sequence.

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.

vided into two categories:

- **Function metrics** assess the pedagogical structure of the dialogue. They are calculated based on the presence, frequency, or sequence of the teaching acts annotated in our dataset (e.g., measuring if a *Rule* is followed by an *Example*). These are detailed in Table 4.

- **Form metrics** evaluate linguistic and stylistic features of the explanation that impact cognitive load and readability, such as lexical complexity or coherence. These are detailed in Table 5.

We investigate this suite through three lenses: human validation, traditional static evaluation, and a novel prompt-based LLM evaluation.

5.1 Human Validation of Metrics

Before applying the metrics, we first sought to validate their relevance with our domain experts. As a follow-up task to the span annotation, we asked our four annotators to assess each of the 13 metrics for every dialogue, with reference to the descriptions provided in Table 4 and Table 5. Using a 3-point Likert scale, they rated the **presence** of each metric and its **contribution** to the explanation's quality for the given knowledge level. This step anchors our framework in the expertise and judgment of education professionals.

The results of the annotators' assessment of metric presence are shown in Figure 5a, based on the normalized average of the ratings. The analysis reveals a strong alignment between the perceived presence of most function metrics and the explainee's knowledge level. For instance, *Check* *for prior knowledge (PK)*, *Rule-example (RE)*, and *Example-rule (ER)* structures are rated as more present in dialogues with less expert explainees. In contrast, form-based metrics like *Adaptation (AD)*, *Readability (RL)*, and *Coherence (CO)* are consistently rated as important across all levels, indicating that they serve as foundational elements of any strong explanation.

5.2 Static vs. Prompt-based Evaluation

We then evaluated the dialogues automatically using two different methods to see how well they could replicate the nuanced judgments of our human experts.

Static Evaluation. Our first approach uses "static" or rule-based calculations. For function metrics, this involves counting the tokens in the corresponding gold-standard teaching act spans (e.g., T01 for PK). For form metrics, we use standard linguistic feature calculations like the Flesch-Kincaid grade level for readability (RL). The results, shown in Figure 5b, reveal a key limitation of this approach. While some form-based metrics (e.g., LC, SD, RL) show a clear trend across knowledge levels, the function-based metrics appear noisy and fail to show a consistent correlation. The static method seems too superficial to capture the functional quality of the instructional discourse.

Prompt-based Evaluation. To overcome these limitations, we developed a "prompt-based" evaluation framework inspired by Rooein et al. (2024). Instead of relying on simple counts, we leverage an LLM's reasoning capabilities. We prompted `GPT-4o` with the full dialogue and asked it to rate each metric on a scale from 0 to 10 (e.g., "On a scale from 0 to 10, how well does the explainer check for understanding?").

The results, shown in Figure 5c for the function metrics, are strikingly different from the static evaluation. The prompt-based scores align remarkably well with the human judgments from our validation step. There is a clear, graded relationship between the metric scores and the explainees' knowledge levels, especially for *PK*, *EA*, *RE*, and *ER*. This demonstrates that an LLM-based "judge" is far more capable of capturing the nuanced, functional aspects of instructional quality than simple static heuristics. For form-related metrics (Appendix F), the prompt-based scores were high and stable across levels, confirming the human assessment that these are universally important. This suggests a hybrid approach for future work: static metrics may suffice for form, but evaluating the functional discourse structure of explanations requires the inferential power of prompt-based LLM evaluation.

6 Discussion

Our findings offer several key implications for the fields of computational discourse analysis, educational technology, as well as NLP practices such as fine-tuning and automated evaluation.

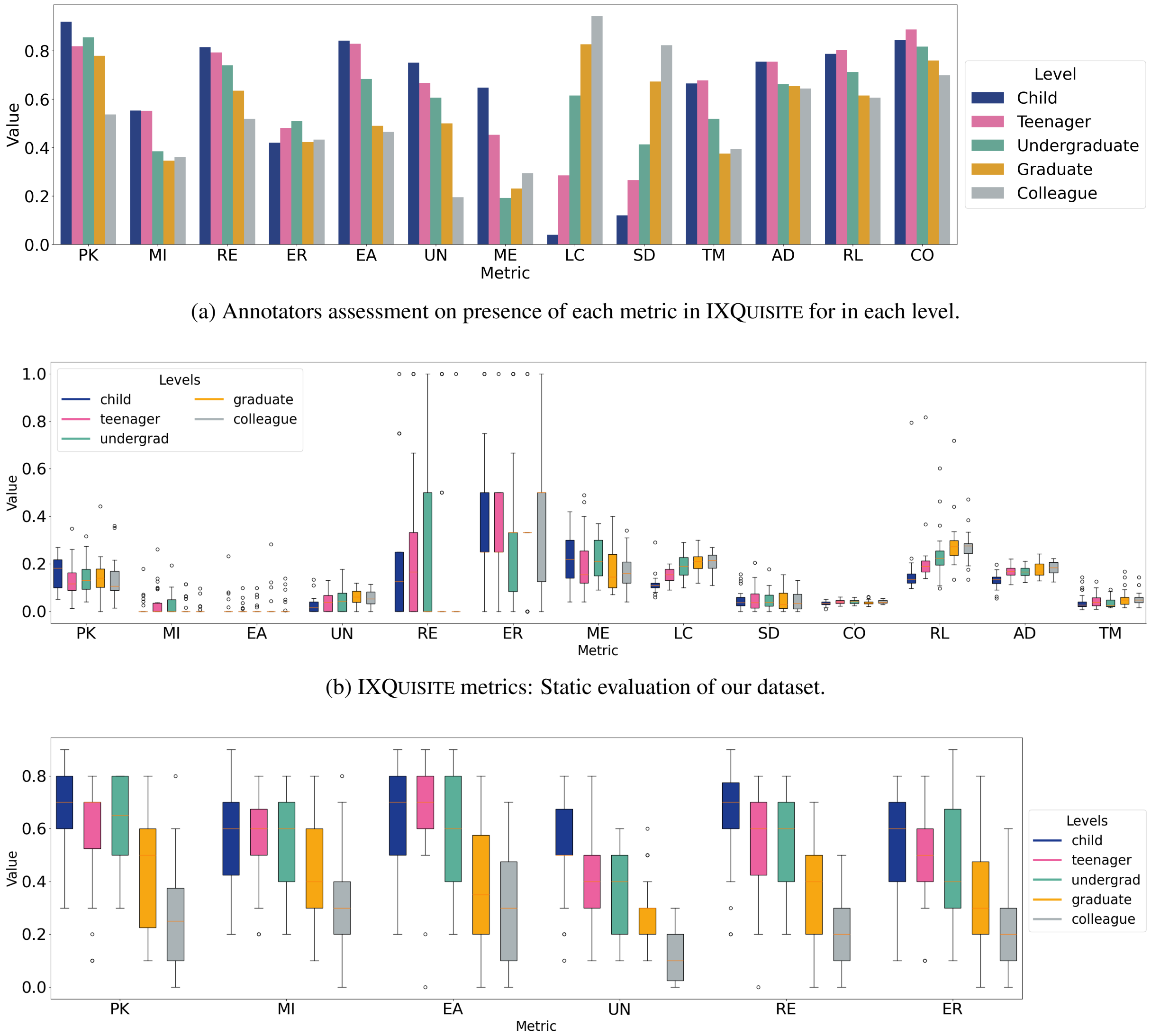
Implications for Discourse Analysis. Our work treats teaching as a complex, goal-oriented discourse phenomenon. By creating a fine-grained, span-level annotation scheme for pedagogical strategies, we provide a new lens for analyzing dialogue structure. The *teaching acts* in ReWIRED can be viewed as a domain-specific set of discourse relations that govern how instructional conversations are built. Our dataset, with its unique five-level structure of explainee expertise, offers a controlled environment to study **audience adaptation** at a granular level. Future work can analyze the typical sequences and flows of these acts to uncover the "discourse grammar" of effective explanation.

Implications for Educational Technology and XAI. Our contributions provide a direct pathway toward more effective and pedagogically-aware AI systems, e.g.:

- **AI Tutors:** An automated tutor could use our models to self-assess its own dialogue strategies in real-time (Wang et al., 2024). If it produces too many 'Knowledge Statement's without a corresponding 'Check for Understanding', it could adapt its strategy to be more interactive. The IXQUISITE metrics could serve as a reward function for RL-based dialogue managers.

- **Tools for Human Educators:** Our framework could power tools that provide feedback to trainee teachers. By analyzing a transcript of a practice lesson, such a tool could highlight strengths (e.g., "Great use of analogy here!") or suggest improvements (e.g., "Consider first checking for prior knowledge.").

- **Advancing Explainable AI (XAI):** True XAI should go beyond presenting information to actively fostering human understanding. Our work offers a blueprint for pedagogically sound explanatory dialogue, shifting the focus



(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.

from producing static explanations to enabling interactive and adaptive exchanges (Feldhus et al., 2023).

Methodological Takeaways for NLP. Finally, our experiments offer two clear methodological lessons. First, for complex, domain-specific structured prediction tasks like identifying teaching acts, **in-domain fine-tuning is critical**. It vastly outperforms even the most capable few-shot LLMs, demonstrating that task-specific adaptation remains essential for high-fidelity discourse analysis. The exceptional performance can be explained with the fact that the ground truth is a consolidation from multiple annotators. The model is exposed to many examples of the already consolidated teaching acts, which is in contrast to how human annotators are typically introduced to labeling efforts, namely with explicit instructions and few-shot examples. This is reinforced by our observation that models exposed only to few-shot examples without fine-tuning performed substantially worse.

Second, our work combines the strengths of two approaches: from the **LLM-as-a-judge** paradigm and static metrics. Our analyses suggest that for evaluating nuanced pragmatic qualities of discourse, leveraging the contextual reasoning of LLMs is a more promising path forward than relying on surface-level heuristics. However, it should be taken into consideration that, depending on the task, judge models' agreement with human annotators can vary across datasets and domains (Bavaresco et al., 2025). In future work, applying the same principles across multiple LLMs may yield different outcomes.

7 Conclusion

In this paper, we introduced ReWIRED, a dataset of instructional dialogues significantly extending prior work with expert, span-level annotations of teaching acts. We demonstrated that while automatically labeling these acts is challenging for few-shot LLMs, fine-tuning achieves excellent performance with both smaller and larger models, establishing a reliable methodology for analyzing pedagogical discourse at scale. Furthermore, we proposed a framework for evaluating the quality of these explanations, showing that while static metrics are limited for certain dimensions, a prompt-based approach using LLMs as evaluators more effectively captures how instructional strategies are adapted to explainees' knowledge levels.

Our contributions provide a crucial bridge between pedagogical theory and computational discourse analysis. The dataset and validated evaluation suite offer a concrete methodology for building and assessing systems that engage in instructional dialogue. This paves the way for a new generation of applications, from more adaptive and effective automated tutors to AI-powered tools that provide feedback to human educators. Ultimately, by modeling the structure of effective teaching, our work helps advance the broader goal of creating AI systems that can not only explain, but explain well.

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, following 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 four 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.

Acknowledgements

We thank Christoph Kulgemeyer and David Buschhüter for their input and advice on didactics, the annotators and data curators, including Kevin Kornetzki-Lucza, Iris Vassiliadi, Marie-Theres Ebner, Sören Wulff, Lavanya Govindaraju, Elif Kara, Christopher Ebert, and Emiliano Valdes Menchaca. Finally, we thank Qianli Wang for providing preprocessing code at an early stage of the project and David Schlangen and Tatiana Anikina for their feedback on earlier drafts of the paper. This work was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant numbers 03RU2U151C (news-polygraph), 01IW23002 (No-IDLE), and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg, and partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 - 438445824.

References

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.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

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.

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.

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.

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.

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.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5399–5421, Singapore. Association for Computational Linguistics.

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, and 55 others. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *CoRR*, abs/2407.12687.

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.

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, and 34 others. 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*.

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.

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 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.

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.

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

Appendix

A Examples for acts

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

B Label distributions

Figure 9 shows the number of distinct acts per dialogue turn as per annotated.

C 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.

D Prompt design

Figure 10 and Figure 11 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 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 12 and Figure 13 show the results from GoLLIE and TANL prompts for `Gemini 1.5 Pro` and `GPT-4o`, respectively.

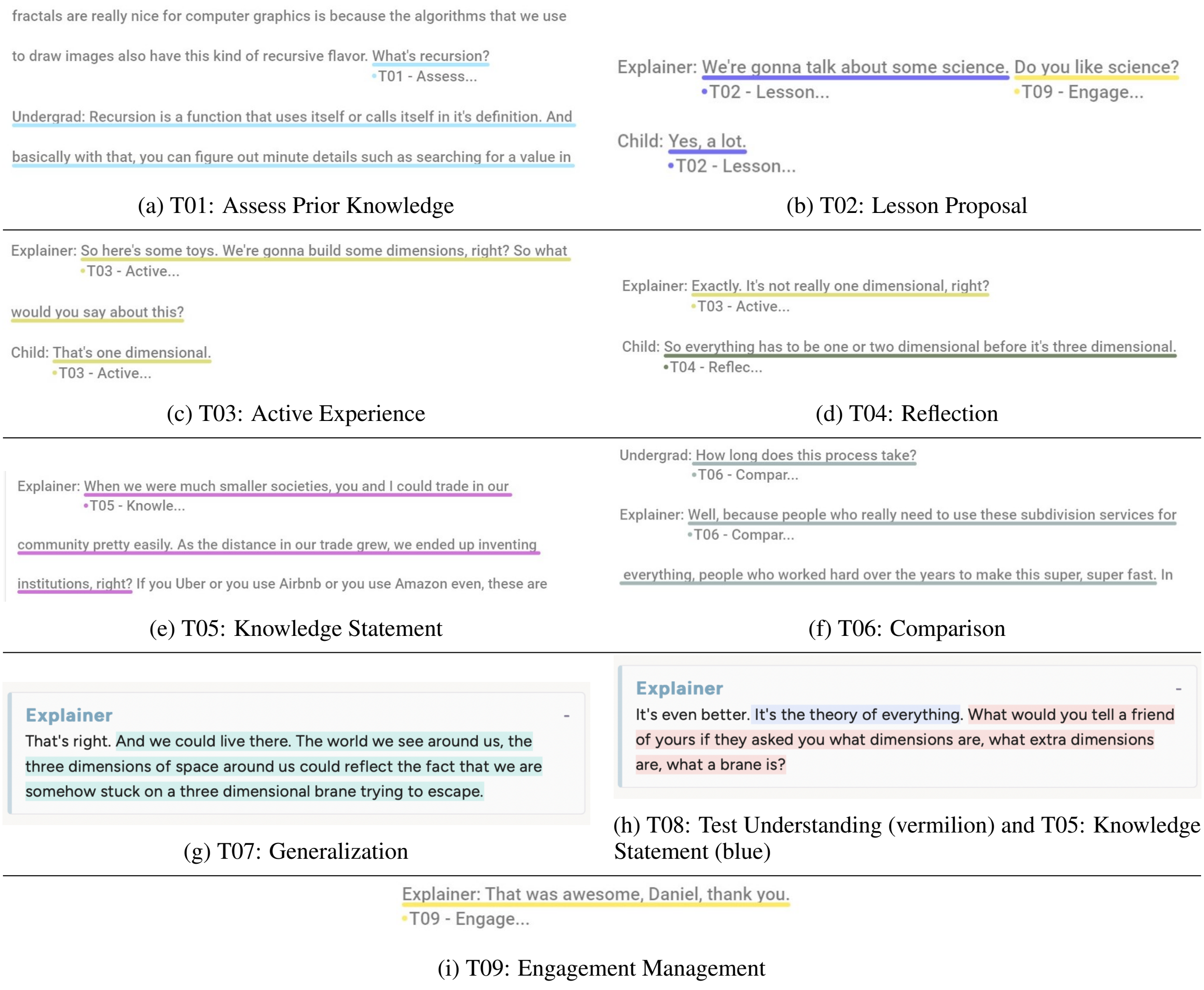
E Annotation instructions

To annotators, we provided examples from Appendix A 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. 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.

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.



fractals are really nice for computer graphics is because the algorithms that we use to draw images also have this kind of recursive flavor. What's recursion?

Undergrad: Recursion is a function that uses itself or calls itself in it's definition. And basically with that, you can figure out minute details such as searching for a value in

(a) T01: Assess Prior Knowledge

Explainer: We're gonna talk about some science. Do you like science?

Child: Yes, a lot.

(b) T02: Lesson Proposal

Explainer: So here's some toys. We're gonna build some dimensions, right? So what would you say about this?

Child: That's one dimensional.

(c) T03: Active Experience

Explainer: Exactly. It's not really one dimensional, right?

Child: So everything has to be one or two dimensional before it's three dimensional.

(d) T04: Reflection

Explainer: When we were much smaller societies, you and I could trade in our community pretty easily. As the distance in our trade grew, we ended up inventing institutions, right? If you Uber or you use Airbnb or you use Amazon even, these are

(e) T05: Knowledge Statement

Undergrad: How long does this process take?

Explainer: Well, because people who really need to use these subdivision services for everything, people who worked hard over the years to make this super, super fast. In

(f) T06: Comparison

Explainer

That's right. And we could live there. The world we see around us, the three dimensions of space around us could reflect the fact that we are somehow stuck on a three dimensional brane trying to escape.

(g) T07: Generalization

Explainer

It's even better. It's the theory of everything. What would you tell a friend of yours if they asked you what dimensions are, what extra dimensions are, what a brane is?

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

Explainer: That was awesome, Daniel, thank you.

(i) T09: Engagement Management

Figure 6: Examples for teaching acts T01-T09.

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 7 shows the annotator's assessment of the importance/contribution of each metric at each level.

F.2 Form metrics: prompt-based evaluation

Figure 8 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.

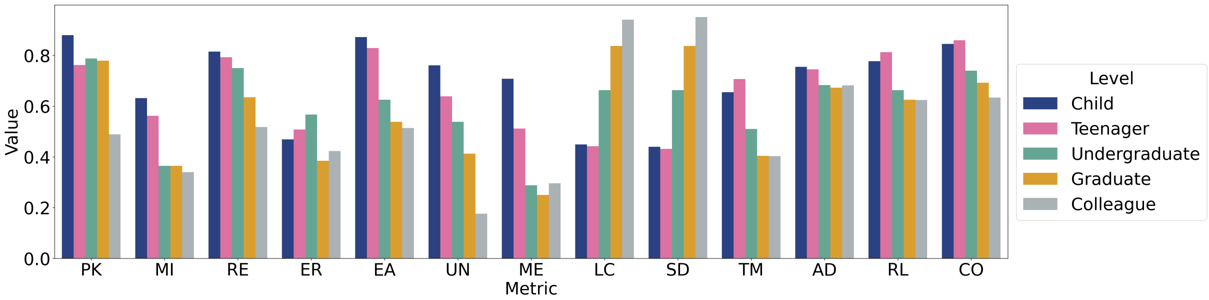


Figure 7: Annotators assessment on contribution of each metric present in IXQUISITE for each level.

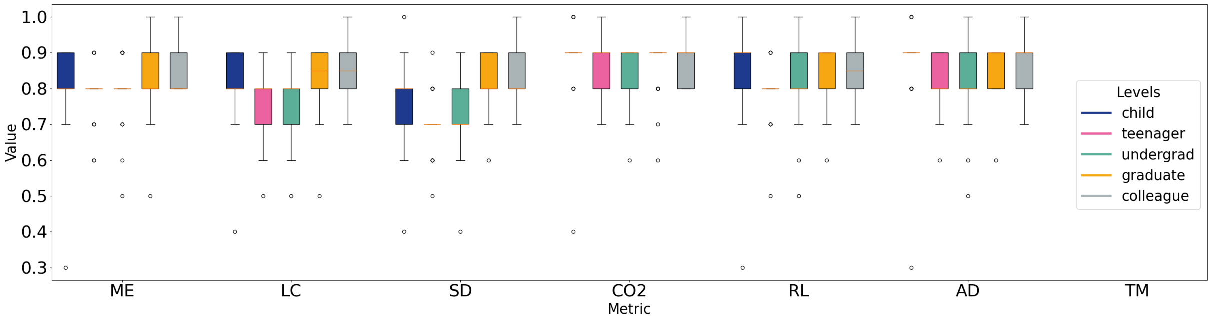


Figure 8: IXQUISITE form metrics: prompt-based evaluation of the five levels in the dataset.

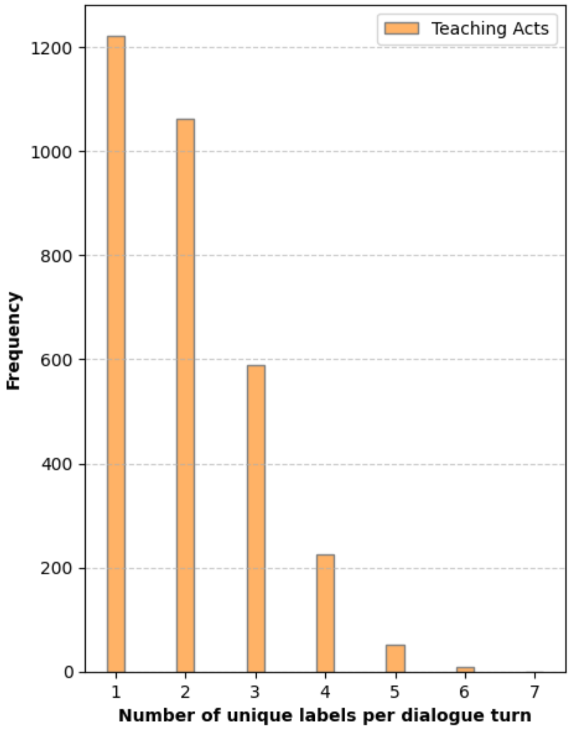


Figure 9: Number of unique teaching acts per turn in ReWIRED. The bar chart reveals that more than half of all dialogue turns in ReWIRED contain more than one distinct teaching act.

```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 10: 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 11: 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 12: Example for a result from a GoLLIE prompt with Gemini 1.5 Pro.

```python
"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 13: Example for a result from a TANL prompt with GPT-4o.