# ReWIRED: Instructional Explanations in Teacher-Student Dialogues

**Anonymous ACL submission**


## Abstract

How to assess the quality of teaching in instructional explanation dialogues is a recurring point of debate in didactics research. For the NLP community, this is a challenging topic thus far, even with the use of LLMs. To address the matter, we create a new annotation scheme of teaching acts aligned with contemporary didactic teaching models. On this basis, we extend an existing dataset of conversational explanations about communicating scientific understanding in teacher-student settings on five levels of the explainee's expertise, with the proposed teaching annotation: explanation and dialogue acts. For better granularity, we reframe the task from a dialogue turn classification to a span labeling task. We then evaluate language models on the labeling of such acts and find that the broad range and structure of the proposed labels is hard to model for LLMs such as `GPT-3.5/-4` via prompting, but a fine-tuned `BERT` can perform both act classification and span labeling well. Finally, we operationalize a series of quality metrics for instructional explanations in the form of a test suite. We find that they match the five expertise levels well and that experts in our data often stick to best practices in teaching.



Figure 1: Instructional explanation dialogue of an expert (center) explaining machine learning to a child (left). Labels on the right indicate the teaching act associated with the turn(s) or span(s) with the same color.

## 1 Introduction

The recent paradigm shift in NLP towards LLMs such as ChatGPT has impacted cross-disciplinary research with education and other social sciences. However, automating teacher coaching (Wang and Demszky, 2023) and student tutoring (Macina et al., 2023) has shown limited success so far. A recent work in tutoring by Lee et al. (2023) has explored creating interactive dialogues to answer children's why and how questions. Measures for estimating the quality of discourse (McNamara et al., 2014) or model-generated explanations (Schuff et al., 2023) exist, but it is unclear how we can assess the quality of teaching in such instructional explanation dialogues and also consider the expertise level of the explainee (Wachsmuth and Alshomary, 2022).

In this work, we first propose a scheme of teaching acts that connect dialogical surface-level utterances with the processes described by two popular teaching models (§2). Thereby, we open the doors to the large-scale analysis of teaching strategies, a goal much sought after in didactics (Matsumura et al., 2008). Secondly, we re-annotate the WIRED dataset from Wachsmuth and Alshomary (2022) to include our scheme of teaching acts, expanding on the two act sets from the original, dialogue acts and explanation acts (§3). The dataset is further enhanced by the inclusion of 45 new conversation transcripts, and by a switch from a turn-labeling to a span-labeling setting for higher granularity.

We evaluate state-of-the-art language models of different sizes on both turn classification and span labeling (§4) and find that large closed-source models cannot perform either task reasonably well and is easily beaten by a fine-tuned BERT. Lastly, to measure "good teaching" according to didactics research in terms of both *meaning* and *form* (Bender and Koller, 2020), we implement a series of quality metrics for instructional explanations, taking into account the presence and order of teaching acts as well as frequency of explanatory patterns. We dub this new test suite IXQUISITE (§5.3) and find that the metrics correlate well with the five expertise levels in our dataset.[1]

With the results and findings of this paper, we contribute to both fields: To NLP, a representation and delineation of teaching acts in one-to-one tutorial dialogues and powerful language models to recognize instructional explanations as well as a sanity check on LLM-based tutoring; to didactics, a new way to look at teaching and lesson-planning at massive scale, by taking a bottom-up approach to modeling the learning and teaching process.

## 2 Background and Related Work

There are many concepts that are common to didactics but are neglected in NLP research. Neither tutoring-related works (Lee et al., 2023; Stasaski et al., 2020) nor concept explanation datasets (Dinan et al., 2019; Jansen et al., 2018) distinguish the type of explanation in social sciences (Miller, 2019) from the interpretation in NLP research.

In science teaching, an explanation is viewed as a practice (or even a purpose) of science or scientists that systematically addresses the questions of "how" and "why" (Kulgemeyer, 2018). Here, **instructional explanations** are those that aim to "communicate a new cognitive model for understanding the world, or how to perform a task, from one understanding-having interlocutor to an understanding lacking one". While most explainability literature has mostly focused on a more philosophical understanding explanation, as that which connects *explanans* and *explanandum* (Miller, 2019), the instructional perspective is closely aligned with the much-needed interest in context for explanations (Mostafazadeh et al., 2020). Despite many systems posing to perform instructional tasks, to our knowledge, they do not take any teaching or learning models into consideration.

**Teaching models** are frameworks to teach teachers how to plan lessons towards better learning outcomes by structuring lessons in accordance with a psychological model of learning. 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 skeptical about the feasibility (Allensworth et al., 2008). The actual instantiation of them in real-world classroom environments is affected by many socio-cultural elements (Ball and Rowan, 2004), which make it hard to evaluate teaching at scale (Matsumura et al., 2008) and objectively, without considering other teaching and social activities surrounding the explanation (Roelle et al., 2015). 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. In this paper, we represent teaching processes (1) in the form of teaching acts (Table 1, Table 5) and investigate if language models can capture the distinctions, and (2) as explanation quality measures (Table 6) and an analysis of how well they correlate with expertise levels of the explainee.

**Tutoring datasets** Our work is closest to Wachsmuth and Alshomary (2022): We re-annotate and extend their dataset, perform similar analyses in terms of statistics and LM experiments, but add a new angle to the data with teaching acts and span-level labeling, allowing us to derive quality events in instructional explanations (§5.3) and experiments with LLMs (§5.2). In contrast to CIMA (Stasaski et al., 2020), TSCC-2 (Caines et al., 2022), and NCTE (Demszky and Hill, 2023), their dataset was a good target for modelling different teaching types, as the varied levels also highlight how teaching can change depending on educational level and course subject. Suresh et al. (2022) and Kupor et al. (2023) both annotated instruction talk moves in classroom settings and their LMs could perform classification tasks well, whereas Macina et al. (2023) and Wang and Demszky (2023) were less successful for applying similar models in neural dialogue tutoring.

**Evaluation of instructional explanations** Previous work in this direction include COH-METRIX, a related suite of measures to assess the quality and readability in discourse automatically (McNamara et al., 2014). Schuff et al. (2023) have also

[1]The dataset, code, and test suite are available at https://anonymous.4open.science/r/ReWIRED/.

| Dialogue acts | Explanation acts | Teaching acts |
|---|---|---|
| **D01**: Check Question<br>Asking a check question | **E01**: Test Understanding<br>Checking whether the listener understood what was being explained | **T01**: Assess Prior Knowledge<br>Checking what the student knows before starting a lesson |
| **D02**: What/How Question<br>Asking a what or a how question | **E02**: Test Prior Knowledge<br>Checking the listener's prior knowledge of the turn's topic | **T02**: Lesson Proposal<br>Proposing the steps that will be taken during the lesson |
| **D03**: Other Question<br>Asking any other question | **E03**: Provide Explanation<br>Explaining any concept or topic to the listener | **T03**: Active Experience<br>Providing the student with puzzle/question to explore; (Student:) Interacting with a mental concept |
| **D04**: Confirming Answer<br>Answering a question with confirmation | **E04**: Request Explanation<br>Requesting any explanation from the listener | **T04**: Reflection<br>Finding gaps in knowledge or inconsistencies; Asking questions about the experience or concept |
| **D05**: Disconfirming Answer<br>Answering a question with disconfirmation | **E05**: Signal Understanding<br>Informing the listener that their last utterance was understood | **T05**: Knowledge Statement<br>Stating the concept(s) being taught via rules or facts |
| **D06**: Other Answer<br>Giving any other answer | **E06**: Signal Non-understanding<br>Informing the listener that the utterance was not understood | **T06**: Comparison<br>Considering similarities and differences between the main concept and other related topics or facts |
| **D07**: Agreeing Statement<br>Conveying agreement on the last utterance of the listener | **E07**: Provide Feedback<br>Responding qualitatively to an utterance by correcting errors | **T07**: Generalization<br>Exploring how the concept applies to new scenarios, experiences and situations outside of the lesson topic |
| **D08**: Disagreeing Statement<br>Conveying disagreement on the last utterance of the listener | **E08**: Provide Assessment<br>Assessing the listener by rephrasing their utterance or giving a hint | **T08**: Test Understanding<br>Finding out if the concept previously established was received correctly and is properly understood |
| **D09**: Informing Statement<br>Providing information with respect to the topic stated in the turn | **E09**: Provide Extraneous Information<br>Giving additional information to foster a complete understanding | **T09**: Engagement Management<br>Maintaining the classroom context to facilitate effective teaching, creating rapport between teacher and student |
| **D10**: Other Act | **E10**: Other Act | **T10**: Other Act |

Table 1: Dialogue, explanation and teaching acts (alongside descriptions) in our ReWIRED dataset.

proposed proxy measures for explanation quality based on syntactic and model-based text generation metrics but found low correlation with human judgments. Demszky et al. (2021) develop a framework for measuring teachers' uptake (defined as *building on the student's contribution via, for example, acknowledgement, repetition or elaboration*). Whitehill and LoCasale-Crouch (2024) explore how LLMs can be used to estimate what they define as "instructional support" domain scores with the help of an observation protocol.

## 3 The ReWIRED Dataset

Wachsmuth and Alshomary (2022) classified parts of instructional explanation dialogues from a dataset collected from the 5-levels video series[2], in which an expert in a topic, such as black holes, or music harmony, explains the topic to people of varying expertise levels:

1. Child,
2. Teenager,
3. Undergraduate college student,
4. Graduate student,
5. Colleague (another expert).

Wachsmuth and Alshomary (2022) introduced two types of conversational acts and used them to model explanation dynamics between explainer and explainee. To increase the models' awareness of teaching perspectives, we add a new scheme of teaching acts to their original two dimensions (Table 1 with supplementary examples in Appendix C) and carry out a refined annotation process. We increase the granularity from a turn labeling to a span labeling task, because the original data did not distinguish between the many moves and intents of an interlocutor, especially in longer turns (Figure 10). We dub this improved dataset ReWIRED. In the following, we will introduce these teaching acts and our annotation process.

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

Table 2: Topics in ReWIRED. 14-22 (yellow) are transcripts that were not part of the original WIRED dataset (Wachsmuth and Alshomary, 2022).

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

| | D01 | D02 | D03 | D04 | D05 | D06 | D07 | D08 | D09 |
|---|---|---|---|---|---|---|---|---|---|
| D01 | 14049 | 515 | 658 | 57 | 12 | 19 | 59 | 14 | 674 |
| D02 | 539 | 13458 | 719 | 4 | 0 | 22 | 45 | 0 | 480 |
| D03 | 719 | 391 | 2976 | 2 | 17 | 24 | 142 | 0 | 955 |
| D04 | 119 | 50 | 2 | 3012 | 110 | 232 | 210 | 12 | 619 |
| D05 | 31 | 17 | 3 | 99 | 1119 | 52 | 66 | 0 | 175 |
| D06 | 0 | 26 | 26 | 207 | 96 | 2400 | 151 | 0 | 143 |
| D07 | 319 | 171 | 98 | 219 | 91 | 182 | 6771 | 4 | 1181 |
| D08 | 10 | 25 | 0 | 94 | 33 | 9 | 153 | 399 | 254 |
| D09 | 978 | 464 | 316 | 676 | 163 | 1071 | 881 | 42 | 216711 |

| | E01 | E02 | E03 | E04 | E05 | E06 | E07 | E08 | E09 |
|---|---|---|---|---|---|---|---|---|---|
| E01 | 2331 | 214 | 508 | 459 | 220 | 5 | 155 | 197 | 28 |
| E02 | 293 | 5760 | 388 | 406 | 64 | 12 | 62 | 170 | 233 |
| E03 | 250 | 384 | 209109 | 1056 | 1325 | 99 | 2980 | 2019 | 665 |
| E04 | 362 | 293 | 854 | 16875 | 150 | 11 | 226 | 469 | 0 |
| E05 | 28 | 117 | 664 | 255 | 7050 | 42 | 673 | 168 | 224 |
| E06 | 0 | 9 | 59 | 99 | 125 | 933 | 26 | 39 | 55 |
| E07 | 30 | 7 | 1087 | 333 | 656 | 32 | 5001 | 619 | 219 |
| E08 | 376 | 441 | 729 | 426 | 34 | 0 | 289 | 4626 | 59 |
| E09 | 49 | 179 | 979 | 262 | 425 | 6 | 458 | 94 | 9165 |

| | T01 | T02 | T03 | T04 | T05 | T06 | T07 | T08 | T09 |
|---|---|---|---|---|---|---|---|---|---|
| T01 | 9240 | 142 | 1365 | 1104 | 909 | 832 | 268 | 101 | 440 |
| T02 | 257 | 2439 | 657 | 188 | 500 | 348 | 307 | 87 | 102 |
| T03 | 513 | 201 | 36459 | 2238 | 8371 | 4279 | 992 | 150 | 852 |
| T04 | 772 | 255 | 6429 | 11865 | 8829 | 3601 | 2508 | 737 | 1377 |
| T05 | 458 | 95 | 1562 | 1065 | 49152 | 3423 | 841 | 157 | 166 |
| T06 | 332 | 24 | 1749 | 474 | 2792 | 5127 | 1302 | 27 | 174 |
| T07 | 122 | 109 | 2773 | 662 | 4249 | 2065 | 10569 | 121 | 607 |
| T08 | 110 | 60 | 1143 | 771 | 679 | 716 | 219 | 3594 | 128 |
| T09 | 154 | 94 | 953 | 530 | 1360 | 396 | 309 | 93 | 6825 |

Figure 2: ReWIRED inter-annotator agreements for the three dimensions dialogue (left), explanation (center) and teaching (right) on token level. For better visibility, we have scale-adjusted 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.

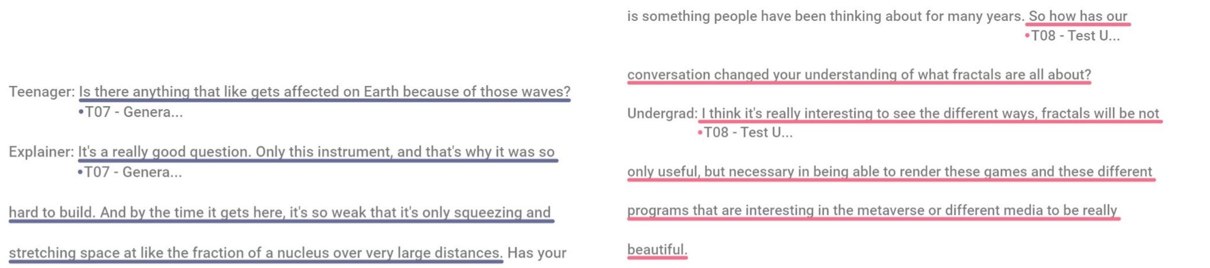



Figure 3: Examples for teaching acts T07 (Generalization) and T08 (Test Understanding).

### 3.1 Teaching acts

Expanding on Wachsmuth and Alshomary (2022), we present a scheme of teaching acts with which to classify dialogues in instructional settings that are coherent with three current and well-accepted teaching models (§2): Teaching as problem solving (**PS**), teaching as concept building (**CB**) (Krabbe et al., 2015), and Oser and Baeriswyl's unified teaching choreographies (**UT**). This is in line with prior work modeling discourse structure in explanations (Bourse and Saint-Dizier, 2012). Concretely, the acts are described in Table 1. Their connection to teaching models and an example[3] are as follows:

- **T01**: *Assess Prior Knowledge* (CB, UT).
- **T02**: *Lesson Proposal* (UT).
- **T03**: *Active Experience* (CB, UT).
- **T04**: *Reflection* (PS).
- **T05**: *Knowledge Statement* (PS).
- **T06**: *Comparison* (UT).
- **T07**: *Generalization* (CB, PS), e.g. Figure 3.
- **T08**: *Test Understanding* (CB), e.g. Figure 3.
- **T09**: *Engagement Management*.
- **T10**: *Other Act*: Any other act that does not fit the above nine acts should instead be placed here.

The main goal of the acts is to bring processes from teaching models closer to the product of their instantiation in actual dialogue (Stolcke et al., 2000), in a way that parts of the dialogue serve as reasonable evidence that the deep processes predicted by teaching models indeed take place.[4]

### 3.2 Annotation

For our annotation task, we asked nine in-house researchers from a (computational) linguistics background to participate in our annotation study. The total of 110 transcripts from 22 topics across five expertise levels (Table 2) were separated into three groups, such that every annotator group annotated the entire dataset exactly once, one third for dialogue acts, one third for explanation acts (using the original act description by Wachsmuth and Alshomary, 2022), and finally one third for our new set of teaching acts. Through three sets of annotations, we aim to reduce the possibility of bias, as some acts are very similar and annotators might be tempted to just repeat previous annotations. For our annotation platform, we used DocCANO (Nakayama et al., 2018), which alleviated the span-labeling task. We additionally randomized all conversations to reduce bias further.[5]

Our inter-annotator agreements are at Fleiss' $\kappa = 0.83$ (dialogue acts), 0.79 (explanation acts)

[3]Acts with a colored border have an example in both Figure 1 and Figure 8.

[4]We consulted three senior didacticians to devise the label scheme. Further details are in Appendix A.

[5]Details on the instructions are provided in Appendix B.

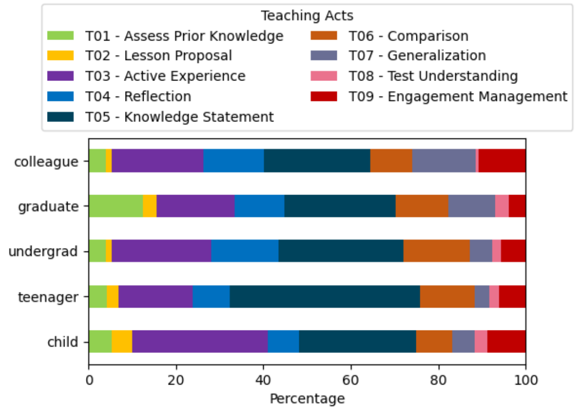



Figure 4: Distribution of teaching acts in ReWIRED across the five expertise levels. Dialogue and explanation act distributions are visualized in Appendix D.

and 0.46 (teaching acts). We plot the nine main labels of each annotation dimension in Figure 2. They show that there also is quite a bit of uncertainty and confusion regarding our teaching acts because our annotators are knowledgeable in computational linguistics but not so much in pedagogy and didactics. Often confused are E03 and E09, as there is a fine line between what we can deem part of an explanation and what is rather supplementary information, and T06 and T07, since both are about "zooming out" of the topic in question and making a broader set of connections to it. The results of our annotation process are visualized via the distribution of teaching acts in Figure 4.

Our annotation scheme differs from DAMSL (Core and Allen, 1997) and ISO 24617-2 (Bunt et al., 2012) in granularity: While there are no dependency relations allowing link structures as in the latter, ours enables finer annotation of semantics related to teaching models. Suresh et al. (2022) presented acts for group dynamics in classrooms related to intents, but misses out on explanation traits and the semantics and pragmatics of the content. Most similar to ours is the CMA schema by Del-Bosque-Trevino et al. (2021) for one-to-one tutorial dialogue sessions: In terms of labels, it vaguely mirrors a lot of acts across all three dimensions, but conflates crucial acts (e.g., *FIM* can be either T02 or T09) and ignores teaching-related concepts. Our acts are, by nature, not as easily recognizable from a surface level due to processes that happen inside the minds of teachers and students.

# 4 Experiments

To evaluate language models on detecting acts across act dimensions, we conduct two experiments: One on turn-level classification, reproducing Wachsmuth and Alshomary (2022), and one on span-labeling for ReWIRED. For both, we test the hypothesis that fine-tuning a masked LM is more consistent at assigning labels on token-level than LLMs prompted for JSON responses indicating spans and labels. We follow Wachsmuth and Alshomary (2022) and evaluate the masked LMs with 5-fold cross-validation, since the number of transcripts is not large enough to define partitions. We provide details on the models in Appendix E.

**Classifying acts** For the turn-level classification of dialogue and explanation acts provided by the original WIRED data, we choose the following baselines: `SVM` with linear kernel for multi-class classification based on `MiniLM` sentence embeddings (Reimers and Gurevych, 2019), and the top-performing `BERT` from Wachsmuth and Alshomary (2022). We compare the following LMs: `BERT` for turn-level classification; `Stable Beluga 2` (`SB2`) (Mahan et al., 2023; Mukherjee et al., 2023), a type of `Llama-2` model (Touvron et al., 2023); `GPT-3.5-turbo-0613`.

**Sequence-labeling acts** For the token-level span labeling task of the three annotation dimensions (Table 1) in our new ReWIRED dataset, we analyze the capabilities of the following LMs: As a baseline, a `BERT` for token-level classification. We compare it to three prompt-based LLMs: `Stable Beluga 2`; `GPT-3.5-turbo-0613`; `GPT-4-0125-preview`. We provide details on the prompt design for the latter three in Appendix F.

# 5 Results and Discussion

## 5.1 Classifying acts

We show the best performance we were able to attain in automatic act classification for all three acts using several LLMs, and compare our results with the results of Wachsmuth and Alshomary (2022).

Table 3 shows that LLMs perform poorly in turn-level dialogue act classification, except for capturing disagreeing statements and answers (D08, D05). The fine-tuned BERT model outperforms all other approaches by a substantial amount. This is also repeated for the explanation act classification: LLMs only excel in recognizing signals of (non-)understanding. Across all sets of classes, however, we also find that none of the approaches is able to capture the labels with a very low amount of data points (D05, D08, E01, T02; see Tables 4 & 9).

| Dialogue acts | D01 | D02 | D03 | D04 | D05 | D06 | D07 | D08 | D09 | D10 | Macro-$F_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| W&A BERT-seq | 76.00 % | 72.00 % | 0.00 % | **35.00** % | 67.00 % | 0.00 % | 69.00 % | 0.00 % | 87.00 % | 61.00 % | 47.00 % |
| SVM + SentTf | 64.30 % | 59.55 % | 0.00 % | 7.14 % | 86.96 % | **7.69** % | 76.28 % | 0.00 % | 83.30 % | 68.57 % | 68.71 % |
| BERT | **87.35** % | **82.81** % | 0.00 % | 0.00 % | 80.77 % | 0.00 % | **82.04** % | 0.00 % | **94.62** % | **76.77** % | **81.67** % |
| SB2 | 20.00 % | 41.51 % | 0.00 % | 14.29 % | **100.00** % | 0.00 % | 28.57 % | 0.00 % | 78.67 % | 0.00 % | 31.45 % |
| GPT-3.5 | 14.33 % | 43.36 % | **4.41** % | 19.15 % | 37.93 % | 5.92 % | 21.41 % | **8.00** % | 69.51 % | 33.88 % | 25.79 % |
| **Expl. acts** | **E01** | **E02** | **E03** | **E04** | **E05** | **E06** | **E07** | **E08** | **E09** | **E10** | **Macro-$F_1$** |
| W&A BERT-seq | **27.00** % | 64.00 % | 84.00 % | 64.00 % | 33.00 % | 21.00 % | 60.00 % | **15.00** % | 8.00 % | 56.00 % | 43.00 % |
| SVM + SentTf | 6.90 % | 66.34 % | 81.37 % | 37.89 % | 13.84 % | 0.00 % | 72.99 % | 0.00 % | **28.07** % | 55.81 % | 63.23 % |
| BERT | 0.00 % | **73.05** % | **93.71** % | **78.26** % | 5.52 % | 0.00 % | **74.89** % | 0.00 % | 0.00 % | **66.04** % | **66.67** % |
| SB2 | 13.79 % | 46.60 % | 81.63 % | 48.89 % | **43.53** % | 18.18 % | 15.13 % | 0.00 % | 9.68 % | 0.00 % | 27.74 % |
| GPT-3.5 | 16.87 % | 38.76 % | 71.70 % | 23.30 % | 37.00 % | **28.30** % | 5.06 % | 0.00 % | 2.86 % | 27.85 % | 27.17 % |

Table 3: Language models evaluated on the tasks of classifying dialogue and explanation acts of whole dialogue turns from the WIRED dataset. We use the previous metrics (W&A BERT-seq) found by Wachsmuth and Alshomary (2022) as our baseline. Percentages under each of the acts show micro-$F_1$ scores.

| Dialogue acts | D01 | D02 | D03 | D04 | D05 | D06 | D07 | D08 | D09 | Macro-$F_1$ | Span Al. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BERT | **73.14** % | **72.72** % | **74.02** % | **55.43** % | **50.25** % | **66.28** % | **60.59** % | **43.14** % | **94.86** % | **69.01** % | – |
| SB2 | 21.66 % | 54.27 % | 2.83 % | 7.63 % | 39.16 % | 9.03 % | 33.66 % | 22.78 % | 93.50 % | 28.72 % | 59.61 % |
| GPT-3.5 | 19.71 % | 54.73 % | 11.69 % | 0.00 % | 8.70 % | 7.01 % | 19.74 % | 12.98 % | 83.87 % | 22.30 % | 59.41 % |
| GPT-4 * | 53.30 % | 51.52 % | 8.34 % | 19.27 % | 33.51 % | 8.54 % | 24.15 % | 19.06 % | 92.65 % | 33.97 % | **63.86** % |
| **Expl. acts** | **E01** | **E02** | **E03** | **E04** | **E05** | **E06** | **E07** | **E08** | **E09** | **Macro-$F_1$** | **Span Al.** |
| BERT | **64.66** % | **67.21** % | **94.69** % | **72.81** % | **64.80** % | **69.09** % | **64.99** % | **80.65** % | **80.34** % | **75.89** % | – |
| SB2 | 8.93 % | 33.63 % | 89.08 % | 56.00 % | 31.67 % | 17.97 % | 20.21 % | 0.00 % | 4.64 % | 26.22 % | 60.54 % |
| GPT-3.5 | 20.06 % | 10.02 % | 84.27 % | 24.23 % | 16.90 % | 19.35 % | 4.69 % | 0.00 % | 7.07 % | 18.66 % | 49.72 % |
| GPT-4 | 27.70 % | 42.11 % | 86.18 % | 66.52 % | 34.82 % | 42.93 % | 19.94 % | 9.07 % | 20.77 % | 35.00 % | **61.49** % |
| **Teaching acts** | **T01** | **T02** | **T03** | **T04** | **T05** | **T06** | **T07** | **T08** | **T09** | **Macro-$F_1$** | **Span Al.** |
| BERT | **81.57** % | **62.38** % | **85.00** % | **80.85** % | **89.61** % | **86.34** % | **85.67** % | **79.57** % | **72.91** % | **82.36** % | – |
| SB2 | 28.24 % | 28.04 % | 13.23 % | 8.42 % | 49.12 % | 7.83 % | 2.09 % | 10.21 % | 29.44 % | 19.62 % | **44.39** % |
| GPT-3.5 | 22.89 % | 8.95 % | 19.10 % | 7.25 % | 40.31 % | 10.31 % | 11.80 % | 5.13 % | 13.66 % | 15.49 % | 31.55 % |
| GPT-4 * | 35.01 % | 26.43 % | 30.06 % | 12.27 % | 43.59 % | 12.77 % | 16.62 % | 11.78 % | 32.51 % | 24.56 % | 39.95 % |

Table 4: Language models evaluated on the tasks of sequence-labeling dialogue, explanation and teaching acts within dialogue turns from our ReWIRED dataset. Percentages under each of the acts show micro-$F_1$ scores. Act 10 was disregarded due to low number of instances, close-to-zero scores and irrelevance for the overall performance. Span Alignment (last column) refers to how well the spans extracted by LLMs align with human-annotated spans. * = Prompting with few-shot demonstrations ($k = 3$) and extended label descriptions.

### 5.2 Sequence-labeling acts

Our results for span-level act prediction (Table 4) reveal that this task is very challenging for the LLMs, since they were not fine-tuned on the task. Still, they can handle the majority classes reasonably well (D02, E04, T05) or very well (D09, E03). However, in all other cases, all LLMs fail to assign the correct label consistently enough. Between the models, GPT-4 has a slight edge over SB2, which in turn is a lot more accurate than GPT-3.5. The difference in model performance is more pronounced for the already established acts (dialogue, explanation), but less so for our new teaching acts, whose label taxonomy is unlikely part of their training data. Evaluating how well the extracted spans align with human-annotated spans (rightmost column) reveals a similar pattern, i.e. GPT-4 beating the rest, except SB2 coming out on top for the teaching acts.

The prompt design that elicits structured prediction in the form of JSON objects from LLMs causes major problems for post-processing. After rigorously handling edge cases, we still find that 12.82 % of SB2, 9.73% of GPT-3.5 and 3.18 % of GPT-4 outputs result in invalid, unparseable JSONs. This can 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 were more consistently structured (1.66% invalid JSONs by GPT-4) and could achieve a noticeably higher performance on the TA task (Zero-shot on TAs: 21.60% Macro-$F_1$), but less so for EA (3-shot on EAs: 33.97%). These findings reflect challenges reported by concurrent related work applying LLMs to dialogue-related tasks (Zhao et al., 2023) and span-labeling tasks (Ziems et al., 2024; Wang et al., 2023) and the general difficulty of applying them to teaching settings (Wang and Demszky, 2023; Macina et al., 2023).

BERT, on the other hand, easily outperformed the prompt-based LLMs across every single act. The stark difference can be attributed to the importance of fine-tuning and the constraint to predict one of the ten acts. For span-labeling tasks such as teaching act classification, we recommend practitioners to employ a controlled setup instead of prompting.

| Category | Description | Origin | Measure |
|---|---|---|---|
| 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 |
| Mindfulness of common misconceptions | The teacher addresses common misconceptions | Wittwer et al. (2010), Andrews et al. (2011) | T04 |
| Rule-Example structure | The teacher states the abstract form of the concept being taught. Then the teacher gives some example to assist understanding | Tomlinson and Hunt (1971) | T05 $\rightarrow$ T03 |
| Example-Rule structure | For procedural knowledge, the teacher first provides examples and then derives the general rule from them | Champagne et al. (1982) | T03 $\rightarrow$ T05 |
| 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 |
| Check for understanding | The teacher tests the understanding of the student | Webb et al. (1995) | T08; E01 |
| Remedial explanations | Either the teacher praises correct understanding (positive reinforcement) or corrects improper understanding | Roelle et al. (2014), Sánchez et al. (2009) | E08 |

Table 5: Explanation and teaching acts-related measures in IXQUISITE for instructional explanation quality based on occurrences of classes from our annotation schema.

| Category | Description | Origin | Measure |
|---|---|---|---|
| Minimal explanations | Low cognitive load, e.g. avoid redundancies (verbosity) such as introducing named entities | Black et al. (1986) | Frequency of named entities |
| 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 |
| 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 |
| 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 |
| Adaptation | The teacher incorporates prior knowledge, misconceptions and interests and uses analogies | Wittwer et al. (2010) | Inverse frequency of synonyms in the text |
| Readability level | Indicator of how difficult a passage is to understand | Crossley et al. (2017) | Flesch-Kincaid Grade level |
| 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 6: Categories for instructional explanation quality and associated numerical measures in IXQUISITE.

### 5.3 Quality Events in Instructional Explanations

Based on our annotation schema and as an additional analysis, we develop and propose a test suite based on didactics research. This novel assessment framework, which is termed as IXQUISITE, addresses both the *form* of instructional explanations (in terms of syntax, vocabulary, etc) and their *function* (as present in the form of different classes in our annotation). While we only carry out analyses on and evaluate the ReWIRED dataset, we are confident that IXQUISITE can be applied to other kinds of instructional explanations, both human- and LLM-generated, among others.

**The IXQUISITE test suite** Since teaching models propose themselves as a proper method for instantiating learning, evaluating teaching according to their adherence to the prescribed method is also natural. We find that teaching models can serve as a quality metric and an opportunity to operationalize many other proposed evaluation metrics from didactics. We provide a new way to interact with the problem by providing a suite of tools that measure quality based on a large selection of proposed quality features from didactics literature. Through our suite of low-level quality tests, we aim to verify didactics theory in a controlled environment at a relatively low cost (using existing libraries, e.g., NLTK, SPACY, and TEXTSTAT). Following the literature review by Kulgemeyer (2018), we track a list of seven events, which, when detected, have been shown to correlate to better learning outcomes, and seven more numerical metrics, which are the discrete values resulting from properties associated with better learning outcome. The events and metrics, along with their descriptions, are listed in Table 5 and Table 6, respectively.

**IXQUISITE results** The qualitative act-based measures, as well as the metrics correlate well with the expert levels present in the ReWIRED dataset

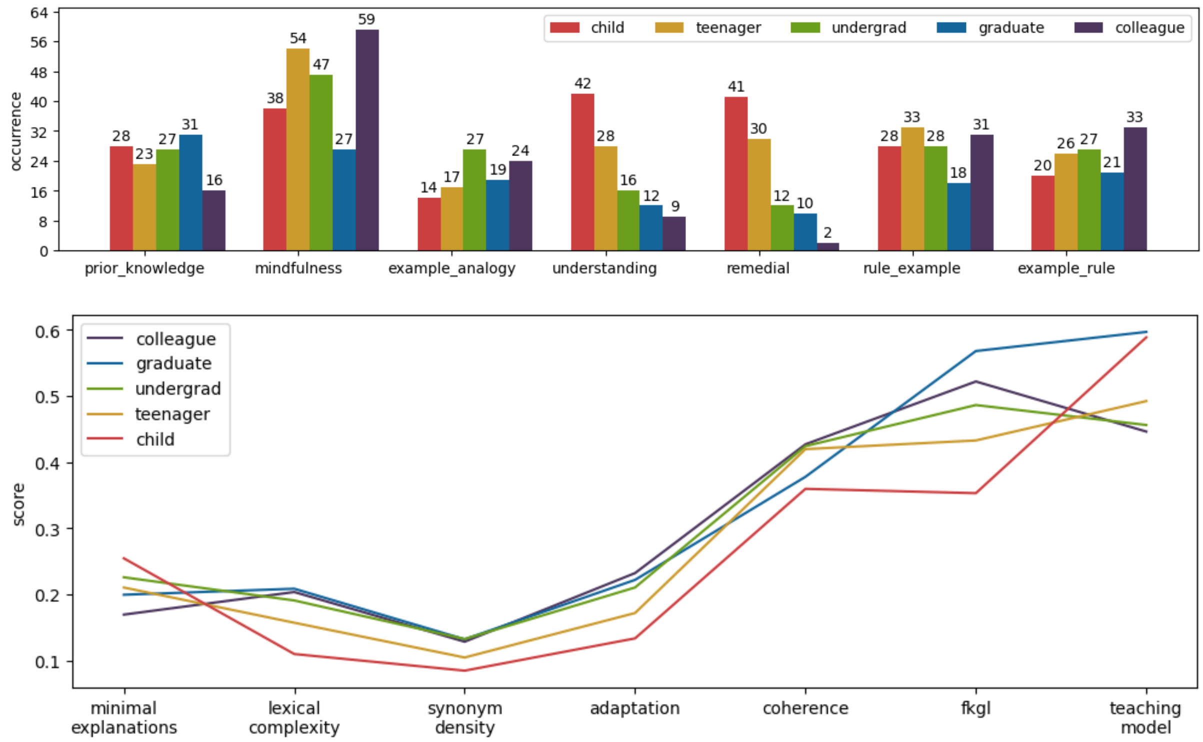



Figure 5: IXQUISITE results with scores from explanation and teaching act-related measures (Table 5; top) and for the five levels in ReWIRED by category according to Table 6 (bottom).

(Figure 5). In terms of the former, testing for understanding and remedial explanations are mostly present in lower expertise levels, which is expected. *Mindfulness (of common misconceptions)* is especially high for colleague-level explanations and reflects the variation in conversation topics present in the dataset. Both *rule-example* and *example-rule* structures are exceptionally present as well as in teenager- and colleague-addressed dialogues.

Regarding our numerical metrics, we observe that explanations tailored to a child present a lower bound across all our metrics, including a lower lexical complexity, reading grade, synonym density, and coherence. However, a general trend is that graduate-level explanations score higher than colleague-grade explanations (e.g., teaching model correlation), because they are more focused on actual topic of discussion, while colleague-grade dialogues might also contain chit-chat and other topics, thus not necessarily following a teaching-like approach. In the case of adaptation, graduate-level explanations are an outlier, where the score is surprisingly lower. Lastly, minimal explanations' scores for children average higher, possibly because of an attempt to establish a common ground with world knowledge via entities.

## 6 Conclusion

We presented an extended dataset of instructional explanation dialogues in one-to-one tutorial sessions, ReWIRED, adding span-level annotations and new teaching acts dimension reflecting good practices according to didactics. Our language model analyses on the span-labeling tasks show that LLMs, including GPT-4, fall far behind controlled setups like a fine-tuned BERT in reliably detecting acts across multiple act dimensions. Our IXQUISITE suite of metrics for quality events in instructional explanations represent the different expertise levels of explainees well and are a first step in operationalizing pedagogical psychological theory for tutorial dialogues in NLP.

In the future, we plan to follow concurrent work in exploring LLM-based explanation quality evaluation (Rooein et al., 2023), especially for metrics such as Adaptation and Mindfulness of common misconceptions. These are hard to capture with the more traditional approach we chose and instead require world knowledge that LLMs can provide. Further data collection and fine-tuning will also allow mimicking the behavior found in classroom transcripts for multi-turn systems. This forms a fertile basis for more satisfactory explanation dialogues from automated tutoring systems.

## Limitations

Resulting from the low inter-annotator agreement for the teaching acts as discussed in §3.2, we want to perform data collection involving teachers and didacticians in the future. However, we point out that even with some teaching acts not being as easily distinguishable as the other act dimensions, our annotators managed to achieve a decent inter-annotator agreement. The single Fleiss' score might be too superficial and that subjectivity and human label variation (Plank, 2022) should be encouraged. ReWIRED includes every single annotator's view and allows a more fine-grained evaluation and countermeasures against "hard labels".

A portion of our test suite relies on human annotation, a factor that may introduce inconsistencies. In this case, replication or extension of the test suite might be difficult without a reliable teaching act prediction model.

Due to time and budget constraints, we were not able to explore many different prompt patterns in our LLM experiments. The prompt design utilized in our study may not represent an ideal formulation, potentially influencing the model's performance.

The dataset we present is extracted from videos - in the transcription, audio and visual elements are not present. 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 any immediate ethical concerns with respect to research and development. The data included in the corpus is readily available from the WIRED web resources. In accordance with 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 nine annotators in our study were paid at least the minimum wage in conformance with the standards of our host institutions' regions. The annotation took each annotator six hours on average, with four at the minimum and twelve at the maximum. In our view, the provided prediction models target dimensions of dialogue turns that are not prone to be misused for ethically doubtful applications.

## References


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*.

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.

Deborah Loewenberg Ball and Brian Rowan. 2004. Introduction: Measuring instruction. *The Elementary School Journal*, 105(1):3–10.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. 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.

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

Sarah Bourse and Patrick Saint-Dizier. 2012. A repository of rules and lexical resources for discourse structure analysis: the case of explanation structures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2778–2785, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

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.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale. *arXiv*, abs/2311.10749.

Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive step-by-step explanatory dialogues to answer children's why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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.

Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. 2023. Stable Beluga models. Hugging Face.

Lindsay Clare Matsumura, Helen E. Garnier, Sharon Cadman Slater, and Melissa D. Boston. 2008. Toward measuring instructional interactions "at-scale". *Educational Assessment*, 13(4):267–300.

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.

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

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical*

*Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv*, abs/2306.02707.

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

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

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

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Roelle, Kirsten Berthold, and Alexander Renkl. 2014. Two instructional aids to optimise processing and learning from instructional explanations. *Instructional Science*, 42:207–228.

Julian Roelle, Claudia Müller, Detlev Roelle, and Kirsten Berthold. 2015. Learning from instructional explanations: Effects of prompts based on the active-constructive-interactive framework. *PLOS ONE*, 10(4):e0124115.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do LLMs adapt to different age and education levels? *arXiv*, abs/2312.02065.

Emilio Sánchez, Héctor García-Rodicio, and Santiago R Acuna. 2009. Are instructional explanations more effective in the context of an impasse? *Instructional Science*, 37:537–563.

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.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

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

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288.

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.

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.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback. *arXiv*, abs/2310.01132v2.

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.

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.

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 Devising the teaching act label scheme

The teaching act scheme has been devised in close collaboration with three senior researchers who are well versed in didactics for physics and math and know the teaching models by heart. The nine main labels are the result of a back-and-forth process spanning several weeks of in-person, virtual meetings, and mails. The process is:

1. Propose a new teaching act that is supported by at least one of the teaching models;
2. Find appropriate examples in both (Re)WIRED and NCTE (Demszky and Hill, 2023);
3. Check potential overlaps with existing labels;
4. Draw clear distinctions between the new label and existing ones.

From the teacher's point-of-view, T03 (*Active Experience*) is about free exploration of the concept or prototype, while T04 (*Reflection*) and T05 (*Knowledge Statement*) are guided comments. From the student's point-of-view, T03 are uncritical and experiential utterances while interacting with the concept, while T04 is the critical highlighting and T06 (*Comparison*) and T07 (*Generalization*) require that a verified concept already exists, usually in the later stages of a dialogue. Although these distinctions were part of the annotation guidelines, Figure 2 (r.) shows that these five labels have the highest disagreement between annotators. We argue that real-world dialogues are messy in this regard and that these gray areas are due to the nature of tutorial dialogues and not a fault of our schema.

In terms of other acts that we considered at the start, we excluded "Experimentation" (including exploratory testing, interaction with test objects, and documentation of observations) from UT and combined it with T03 (*Active Experience*), as this is very much non-verbal and specific to laboratory settings in the natural sciences. From CB, we added "Conceptualization of a prototype" and "Active experience with the concept" to that same act. From PS, we conflated "Understanding a problem" and "Development of solutions" as T05 (*Knowledge Statement*).

Regarding quality assessment, we need to emphasize that the teaching act schema and the quality events in IXQUISITE are not the same. The senior didacticians noted that the perceived quality of the teaching in the NCTE transcripts was poor and made us aware that simply annotating the teaching acts in a dialogue – no matter which data – does not provide us with sufficient signals for how good the teaching is, especially in light of differing expertise levels of the explainee. This is what reassured us that the WIRED dataset with its distinction between five levels was the right one. The lack of quality signals from the simple presence of teaching acts brought us to conceive the IXQUISITE test suite. Table 5 shows that there are many direct correspondences between teaching (and explanation) acts and quality events, but not every act is a signal for teaching quality. That is why we also needed the numerical measures in Table 6 to get a more complete picture of teaching quality.

| Model name | #Params | URL | Training times | Inference times | API costs |
|---|---|---|---|---|---|
| MiniLM | 22.7M | https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 | <1 hour | <1 hour | n.a. |
| BERT | 110M | https://huggingface.co/bert-base-uncased | 13 hours | <1 hour | n.a. |
| Stable Beluga 2 | 70B | https://huggingface.co/petals-team/StableBeluga2 | n.a. | 13 days | n.a. |
| GPT-3.5 & GPT-4 | ? | https://platform.openai.com/docs/api-reference/chat | n.a. | n.a. | $70 |
| GPT-4 3-shot (ReWIRED only) | ? | https://platform.openai.com/docs/api-reference/chat | n.a. | n.a. | $75 |

Table 7: Language models with parameter counts, training times, inference times, and API costs.

## B Annotation instructions

To annotators, we provided examples from Figure 3 and Appendix C as well as further delineations of the acts with examples and descriptions of how to differentiate between them (Appendix A). We also provided a screencast with instructions on how to use DOCCANO 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 DOCCANO 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 larger annotation task is separated into three smaller tasks. It takes around two hours to finish each sub-task.
> We will be trying to label the aforementioned objectives from three different points of view, each with 10 acts: dialogue acts, explanation acts, and teaching acts.
> Dialogue Acts: Focus on basic mechanics in a dialogue between two people
> Explanation Acts: Focus on mechanics of explaining concepts
> Teaching Acts: Focus on conversation mechanics in terms of lesson planning and didactics

## C Examples for acts

Figures 6, 7, and 8 show examples from ReWIRED for each of the acts as provided to the annotators.

## D Label distributions

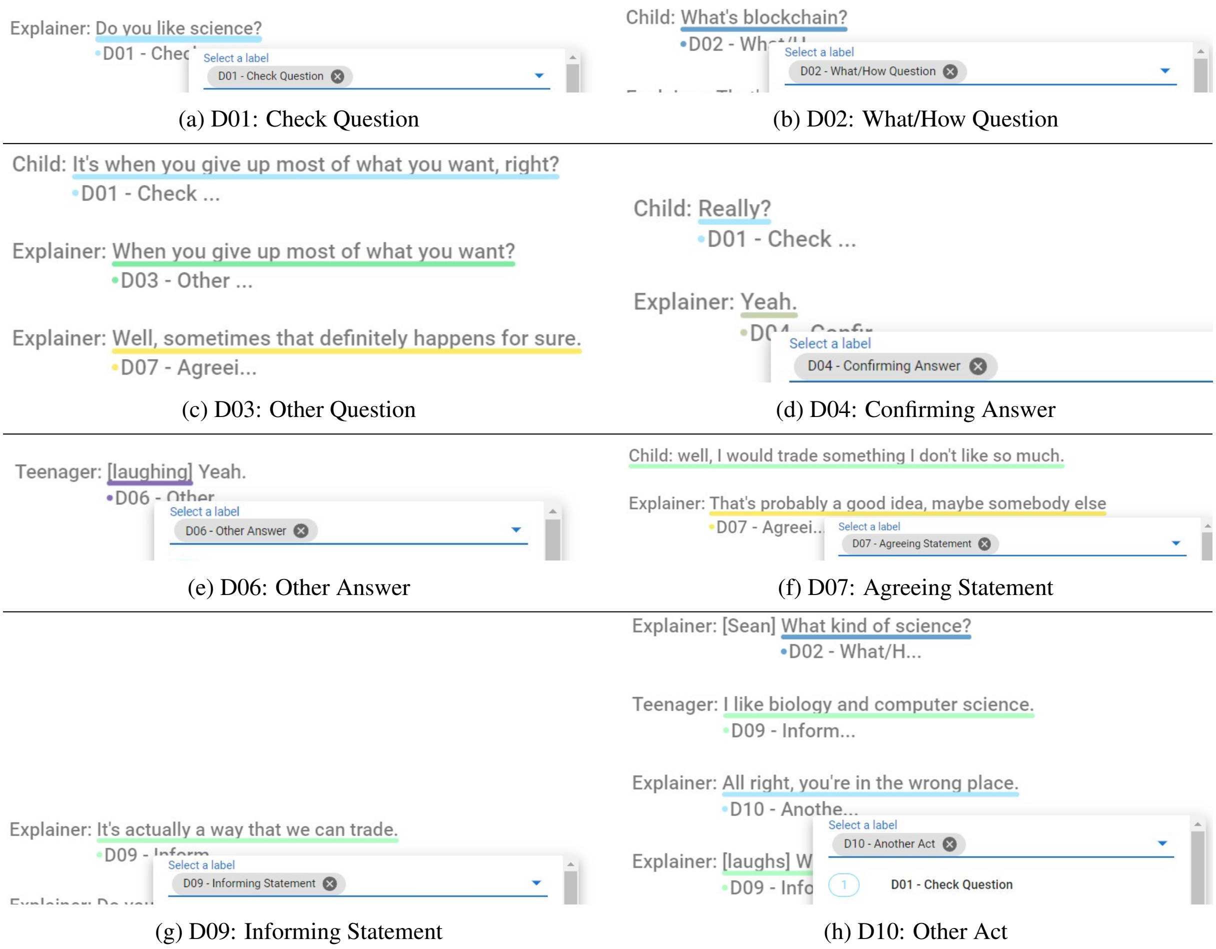
Figure 9 shows the distribution of annotated acts in the dialogue and explanation dimensions. Figure 10 shows the number of distinct acts per dialogue turn.

## E Models

Table 7 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.

## F Prompt design

Figure 11 and Figure 12 depict the prompts used with SB2, GPT-3.5 and GPT-4 to produce the predictions whose evaluation is shown in Table 3 and Table 4, respectively. 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 a JSON format just as we elicit it from the model in the zero-shot setup.




(a) D01: Check Question

(b) D02: What/How Question

(c) D03: Other Question

(d) D04: Confirming Answer

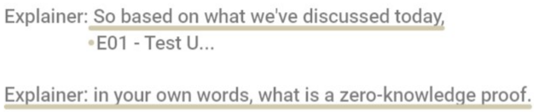
(e) D06: Other Answer

(f) D07: Agreeing Statement

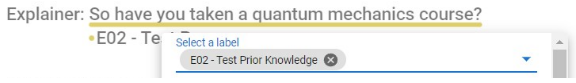
(g) D09: Informing Statement

(h) D10: Other Act

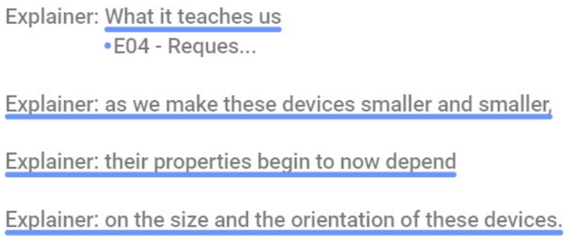
Figure 6: Examples for dialogue acts. D05 and D08 are left out, because they are analogous to D04 and D07, respectively.



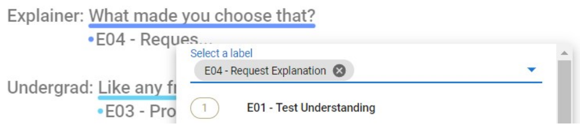

(a) E01: Test Understanding



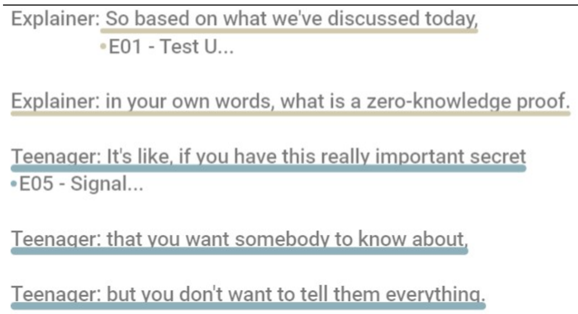

(b) E02: Test Prior Knowledge



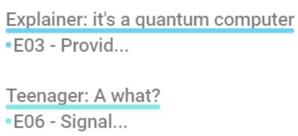

(c) E03: Provide Explanation (The color/label is wrong here!)



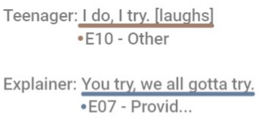

(d) E04: Request Explanation



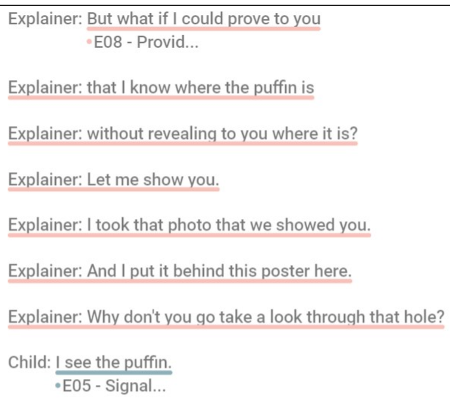

(e) E05: Signal Understanding



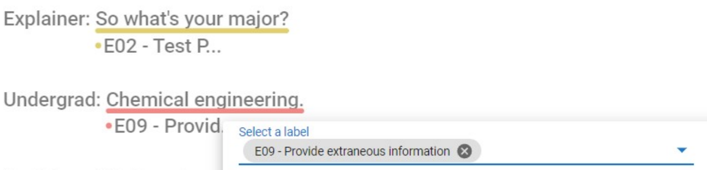

(f) E06: Signal Non-understanding



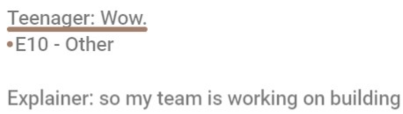

(g) E07: Provide Feedback




(h) E08: Provide Assessment

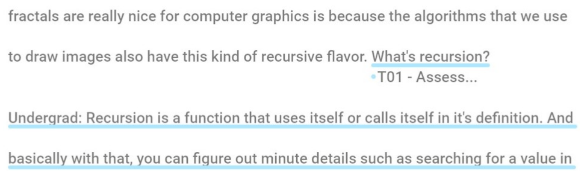



(i) E09: Provide Extraneous Information

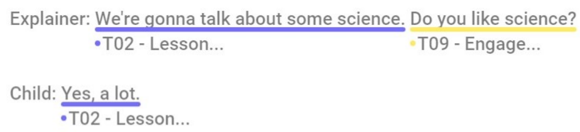



(j) E10: Other Act

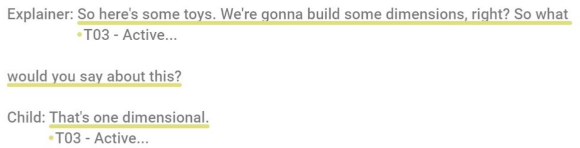
Figure 7: Examples for Explanation Acts.



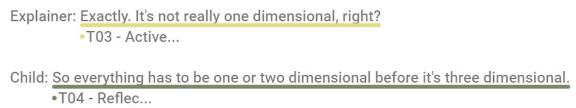

(a) T01: Assess Prior Knowledge



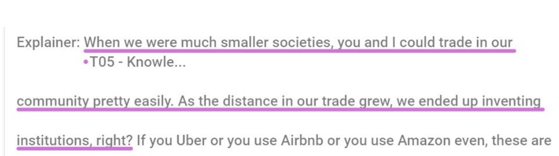

(b) T02: Lesson Proposal



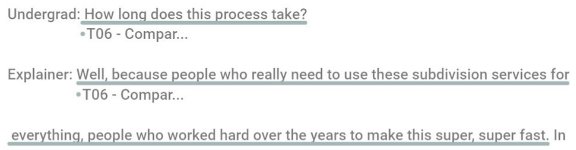

(c) T03: Active Experience




(d) T04: Reflection



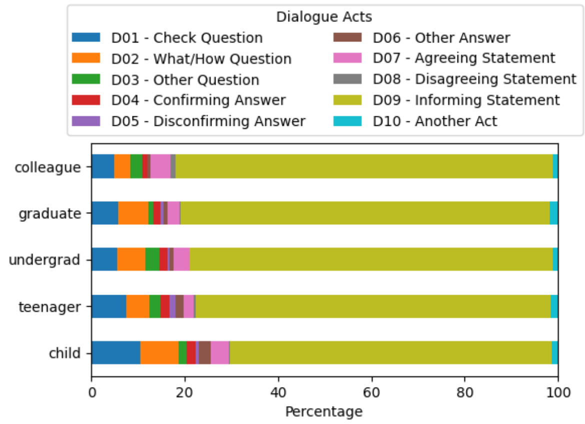

(e) T05: Knowledge Statement



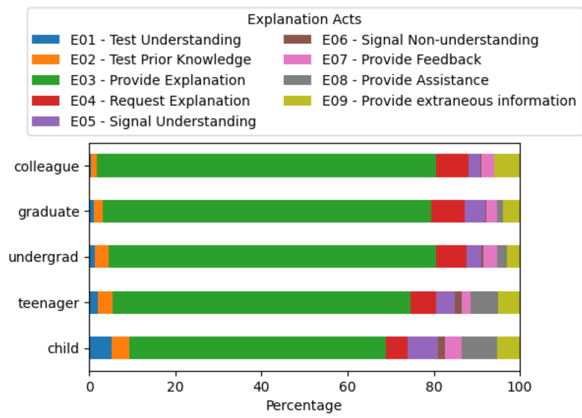

(f) T06: Comparison


(g) T09: Engagement Management

Figure 8: Examples for teaching acts T01-T06 and T09. Examples for T07 and T08 are in Figure 3.




(a) Distribution of dialogue acts




(b) Distribution of explanation acts

Figure 9: Distribution of annotated acts in ReWIRED across the five expertise levels for three dimensions dialogue (a) and explanation (b).

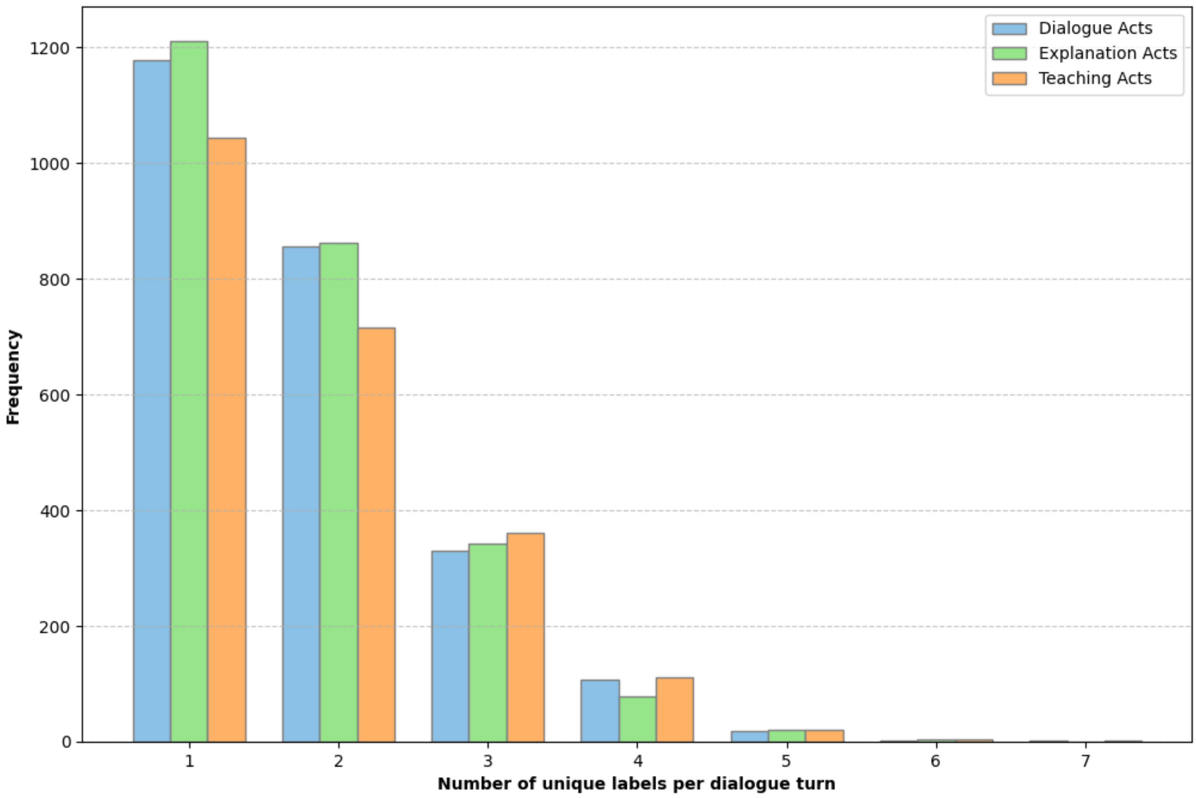



Figure 10: Number of unique dialogue, explanation and teaching acts per turn in ReWIRED. The bar chart shows that more than half of all dialogue turns in ReWIRED contain more than one distinct act, no matter which dimension (dialogue, explanation, teaching) we consider.

```python
system_prompt = (f"You are an expert annotator. In the following, you will be requested to classify a single turn of a dialogue between explainer and {student_role}.\n")
# Example label mapping (dialogue acts)
WIRED_da_label_mapping = {
    '(D01) To ask a check question': 1,
    '(D02) To ask what/how question': 2,
    '(D03) To ask other kind of questions': 3,
    '(D04) To answer a question by confirming': 4,
    '(D05) To answer a question by disconfirming': 5,
    '(D06) To answer - Other': 6,
    '(D07) To provide agreement statement': 7,
    '(D08) To provide disagreement statement': 8,
    '(D09) To provide informing statement': 9,
    '(D10) Other': 10,
}
label_schema = ("The label schema consists of the following 10 classes:\n* " + "\n* ".join(list(WIRED_da_label_mapping.keys())) + "\n")
read_instruction = f"The excerpt from the dialogue:\n{turn_text}\n"
task_instruction = "Predicted label:\n"
# Combine inputs to single string
entire_prompt = system_prompt + label_schema + read_instruction + task_instruction
```

Figure 11: Simplified version of the Python code showing the <u>turn classification</u> task prompt for WIRED.

```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 = ("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")
entire_input = system_prompt + read_instruction + label_schema + task_instruction
```

Figure 12: Simplified version of the Python code showing the <u>span labeling</u> task prompt for ReWIRED.