

# IntrEx: A Dataset for Modeling Engagement in Educational Conversations<sup>1</sup>

Xingwei Tan<sup>1,2</sup>, Mahathi Parvatham<sup>3</sup>, Chiara Gambi<sup>3</sup>, Gabriele Pergola<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Warwick, UK

<sup>2</sup>School of Computer Science, University of Sheffield, UK

<sup>3</sup>Department of Psychology, University of Warwick, UK

Xingwei.Tan@sheffield.ac.uk

{Chiara.Gambi, Gabriele.Pergola.1}@warwick.ac.uk

## Abstract

Engagement and motivation are crucial for second-language acquisition, yet maintaining learner interest in educational conversations remains a challenge. While prior research has explored what makes educational *texts* interesting, still little is known about the linguistic features that drive engagement in *conversations*. To address this gap, we introduce IntrEx<sup>1</sup>, the first large dataset annotated for interestingness and expected interestingness in teacher-student interactions. Built upon the Teacher-Student Chatroom Corpus (TSCC), IntrEx extends prior work by incorporating sequence-level annotations, allowing for the study of engagement beyond isolated turns to capture how interest evolves over extended dialogues. We employ a rigorous annotation process with over 100 second-language learners, using a comparison-based rating approach inspired by reinforcement learning from human feedback (RLHF) to improve agreement. We investigate whether large language models (LLMs) can predict human interestingness judgments. We find that LLMs (7B/8B parameters) fine-tuned on interestingness ratings outperform larger proprietary models like GPT-4o, demonstrating the potential for specialised datasets to model engagement in educational settings. Finally, we analyze how linguistic and cognitive factors, such as *concreteness*, *comprehensibility* (*readability*), and *uptake*, influence engagement in educational dialogues.

## 1 Introduction

Engagement and motivation are fundamental when learning a second language, influencing both learning outcomes and retention (Dörnyei and Ushioda, 2021; Masgoret and Gardner, 2003). However, while existing studies emphasize the relevance of content (Lee and Pulido, 2017; Goris et al., 2019; Tan et al., 2025b), it is still an open research question how linguistic properties themselves shape

<sup>1</sup><https://huggingface.co/collections/XingweiT/intrex-68a8f2c97688157066860ae2>

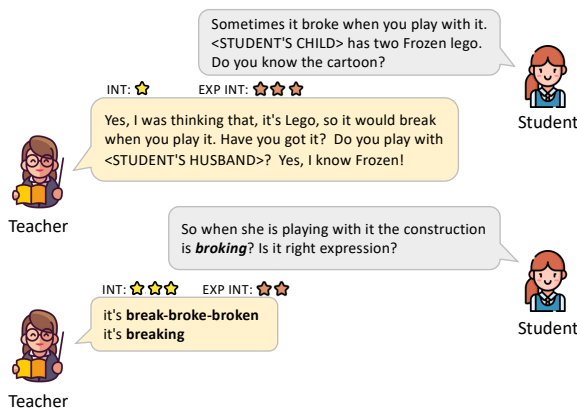


Figure 1: An example of the student-teacher conversations in the IntrEx dataset, along with interestingness (INT) and expected interestingness (EXP INT) scores.

engagement in conversations. This is particularly relevant in educational settings, where structured dialogue plays a central role in knowledge transfer, especially as dialogue-based learning environments continue to expand with the use of AI tutors based mainly on dialogue interactions (Caines et al., 2022; Team, 2023; Tan et al., 2025a).

Existing research has identified linguistic features, such as concreteness, that contribute to the interestingness of texts, often relying on whole-document assessments (Sadoski, 2001; Pergola et al., 2019, 2021; Lupo et al., 2019; Lee and Lee, 2023; Nguyen et al., 2024). Yet, in *educational* conversations, engagement is not static but shaped by thematic continuity, discourse structure, and interaction flow. Whole-document assessments offer limited insight into how interest<sup>2</sup> develops across multiple exchanges, making it essential to analyze sequence-level discourse features.

To investigate this, we build upon the Teacher-Student Chatroom Corpus (TSCC) (Caines et al., 2020, 2022), the largest dataset of chat-based edu-

<sup>2</sup>While we use the terms *interest*, *curiosity* and *motivation* interchangeably, we acknowledge nuanced distinctions between these concepts as discussed in previous literature (Krapp, 1994; Peterson and Hidi, 2019; Donnellan et al., 2022).

cational interactions between teachers and second-language learners. The TSCC provides pedagogical discourse annotations, such as *topic opening* and *explanation*. However, while these annotations describe functional aspects of conversation, they do not account for the engagement and interestingness experienced by learners.

To bridge this gap, we introduce IntrEx, a dataset for modelling interestingness and its expectation in educational conversations, resulting from a comprehensive annotation process, involving over 100 second-language learners, to introduce sequence-level annotations of *interestingness* that, for the first time, capture at scale how interestingness and expectations evolve across extended interactions. We introduce two key dimensions of conversational engagement: *interestingness*, reflecting the perceived level of engagement, and *expected interestingness*, capturing the anticipated level of engagement. These two scores indicate not only what learners found interesting, but also what they expected to find interesting (Murray, 2022). However, assessing interestingness is inherently subjective and prone to variability among annotators. To improve consistency, we thus devised a comparison-based annotation approach inspired by reinforcement learning from human feedback (RLHF) strategies (Stiennon et al., 2020; Ouyang et al., 2022). Instead of rating conversational turns in isolation, annotators compare original utterances with automatically rewritten ‘boring’ versions, leading to higher agreement and reduced noise.

nabled by our new annotations, we conduct two main analyses focused on: (i) the linguistic features that contribute to interestingness, and (ii) the notion of interestingness encoded in LLMs. For linguistic features, we analysed *concreteness*, *comprehensibility*, and *uptake* via a wide range of metrics (Hosseini et al., 2022; Lyu et al., 2024), disentangling how each factor influences engagement in teacher-student interactions using linear mixed-effect regressions. We conduct experiments on standard vs. instruction-based fine-tuning and analyze how our dataset improves alignment with human interest judgments. Notably, we find that fine-tuned models with only 7B–8B parameters outperform larger models like GPT-4o (OpenAI, 2024), in predicting interestingness in educational dialogues, demonstrating the potential of this new dataset for reward-modelling and for fostering research aimed at improving conversational models

for second-language learning. We then conduct a comprehensive investigation of the ability of LLMs to predict *interestingness*, similarly to human annotators. Recent work (e.g., LaMDA (Thoppilan et al., 2022)) has incorporated interestingness as an implicit scoring metric, but no existing dataset provides explicit human-labeled annotations for engagement in educational dialogues.

Our contributions can be summarized as follows:

- We introduce the first dataset annotated for interestingness and expected interestingness in educational conversations for second-language learners. We release individual annotator ratings, demographic information, and aggregated scores to ensure dataset transparency.
- We conduct an experimental assessment exploring how our dataset can facilitate the alignment of LLMs with human interest in learning settings. By fine-tuning with IntrEx data, we show that small LLMs have the potential to outperform larger LLMs in predicting conversational engagement.
- We conduct an extensive analysis of how linguistic and cognitive factors influence engagement in teacher-student dialogues, specifically focusing on concreteness, comprehensibility (readability), and uptake.

## 2 Related Work

### 2.1 Conversational Corpora

Several corpora have been developed to study spoken language in various contexts. The Cambridge and Nottingham Corpus of Discourse in English (CANCODE) records spontaneous conversations in diverse informal settings and has been used to study the grammar of spoken interaction (Carter and McCarthy, 1997). The British National Corpus features transcriptions of spoken conversation captured in settings ranging from parliamentary debates to casual discussion among friends and family (Love et al., 2017). Corpora based on educational interactions, such as lectures and small group discussions, include the widely-used Michigan Corpus of Academic Spoken English (MICASE) (Simpson et al., 2002), the TOEFL2000 Spoken and Written Academic Language corpus (Biber et al., 2004), and the Limerick Belfast Corpus of Academic Spoken English (LIBEL) (O’Keeffe and Walsh, 2012). More specialized corpora include the Why2Atlas

Human-Human Typed Tutoring Corpus (Rosé et al., 2003), containing physics tutoring chats, and a corpus of conversations between native speakers and learners of Japanese collected in a VR campus Toyoda and Harrison (2002). The Teacher-Student Chatroom Corpus (TSCC) (Caines et al., 2020) and its extended version (Caines et al., 2022) contain chatroom dialogues and are annotated with conversational analysis of sequence types, pedagogical focus, and correction of grammatical errors. However, none of the aforementioned corpora provide annotations for *interestingness*, which is the focus of our current work.

## 2.2 Interestingness

Researchers across the fields of psychology and education have extensively investigated the factors associated with *interestingness*, here defined following Thoppilan et al. (2022), as the degree to which a text captures attention or sparks curiosity. Murayama (2022) discussed the neuro-cognitive mechanisms underlying human interest, suggesting that learning, as an information-seeking behaviour, is driven by reward prediction errors: Positive prediction errors (rewards exceeding expectations) motivate further exploration, while negative prediction errors diminish motivation. Tin (2009), controlling for various lecture characteristics, found that tangible, personalized, and contextualized content increases learner interest during lectures. Other prominent psychological theories suggest a complex relationship between interest and comprehensibility or complexity (Wharton, 1988; Dubey and Griffiths, 2020; Oudeyer et al., 2016). The “Goldilocks effect” (Kidd and Hayden, 2015) posits that learners prefer stimuli that are neither too simple nor too complex for their current understanding. This aligns with the information gap theory (Loewenstein, 1994), which proposes that the desire for information is fueled by perceived gaps in existing knowledge. Stimuli that are too simple present minimal knowledge gaps and thus elicit little interest, while overly complex stimuli create insurmountable gaps, leading to avoidance rather than interest. Therefore, comprehensibility likely needs to fall within an optimal range to maximize interest. Texts that are either too easy or too difficult will be less engaging than those at an appropriate level of comprehensibility. This also implies that highly concrete texts, potentially easier to comprehend, may not correlate positively with interest, as they could be perceived as overly

simplicistic and thus less engaging.

## 3 The IntrEx Dataset

IntrEx provides annotations for the conversations in TSCC V2. This section describes TSCC V2 and the IntrEx annotation process. Fig. 2 shows the overall process.

### 3.1 Background: The TSSC V2 Corpus

TSCC V2 (Caines et al., 2020, 2022) comprises the conversation histories of one-to-one English learning lessons conducted in online and private chatrooms. Each lesson typically lasts about an hour. In total, the dataset contains 260 conversations involving 2 teachers and 12 students. A *dialogue snippet* is defined as an exchange of messages, generally consisting of one message (a *turn*) from the teacher and one from the student. The average number of dialogue snippets per conversation is 52 (ranging from 26 to 90). The conversations are manually annotated for sequence type, indicating major or minor shifts in the conversational flow (e.g., *opening*, *exercise*, *redirection*), and teaching focus, which identifies the skills targeted within each sequence.

### 3.2 Task Definition

We annotate two types of engagement scores: *interestingness* and *expected interestingness*. They are both integer values ranging from 0 to 4. Following Thoppilan et al. (2022), we define *Interestingness* as the degree to which the current message captures the annotator’s attention or sparks curiosity. *Expected interestingness* represents the annotator’s prediction of their *interestingness* level after receiving the next message. Since annotators provide this score before seeing the next message, it reflects speculative engagement expectations rather than actual interest. By collecting *expected interestingness*, we can quantify how new topic introductions or discourse shifts exceed, meet, or fall short of expectations. This allows us to analyze the role of anticipation in shaping the potential engagement for the students, offering insights into how conversational structure influences perceived interestingness. The definitions are presented to the annotators as shown in Figure 3 and 4 in Appendix F. Presenting explicit yet intuitive definitions helped reduce potential bias from overly prescriptive instruction, while still allowing for natural inter-individual variability in how engagement is perceived. To account for this subjectivity, each

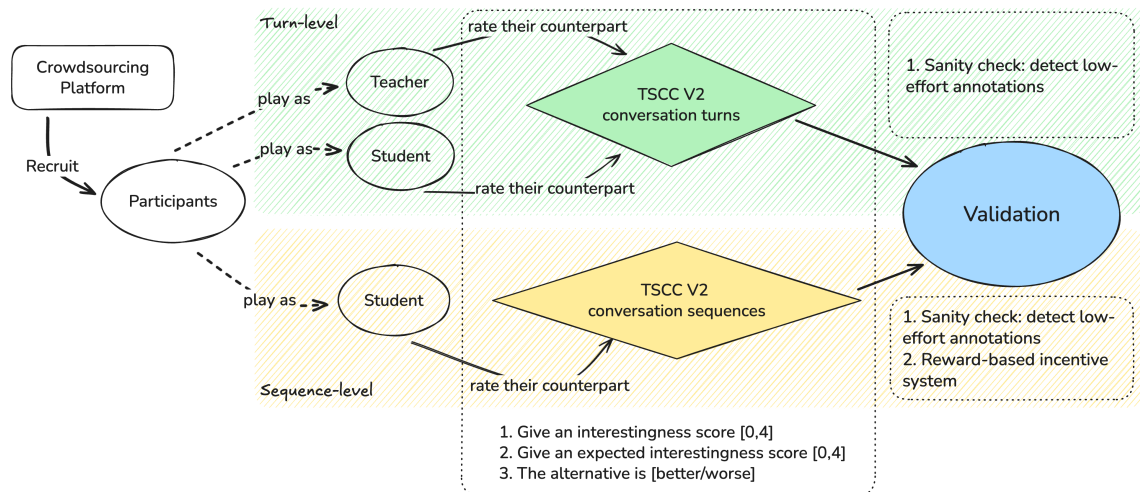


Figure 2: The overall process of the annotation to build IntrEx.

part of the dataset was independently rated by three different annotators, ensuring a diverse range of judgments.

There are two types of annotations in IntrEx: turn-level and sequence-level.

**Turn-level Annotations.** During the turn-level annotation task, half of the annotators imagined themselves as the teacher in the conversation and rated the interestingness of the student’s responses. The other half imagined themselves as the student in the conversation and rated the interestingness of the teacher’s responses. Our annotation platform (App. F, Fig. 5) displays the conversation page by page, where each page contains a dialogue snippet: a teacher turn and a student turn. Annotators were asked to assign an interestingness score and an expected interestingness score based on the turn from the other party. However, we observed that interestingness scores often remained unchanged across consecutive messages, particularly within the same topic. This led to annotator fatigue, reducing annotation quality and lowering inter-annotator agreement. Additionally, engagement in conversations is not always localized to a single turn; it evolves over longer sequences, making turn-level scores too fine-grained to capture conversational dynamics.

**Sequence-level Annotations.** To mitigate this, we segmented each conversation into sequences based on the labels provided by the TSCC V2 dataset. The sequence type labels indicate major conversational shifts, such as a change in teaching contents or in discourse structure. A new sequence is initiated whenever a message’s associated sequence type differs from that of the preceding message. This approach ensures that each sequence repre-

sents distinct segments of the lesson, such as *homework*, *clarification*, or *closing*. Instead of evaluating each turn, annotators provide an *interestingness* and an *expected interestingness* score for a whole sequence, substantially reducing their workload. In this setting, the annotators are only required to rate teachers’ messages in the sequence-level annotation, because (i) teachers deliver significantly more content, (ii) and the sequence type labels are often about teaching content.

### 3.3 Data Collection and Validation

#### Annotation Design and Recruitment Strategy

Participants were recruited via Prolific<sup>3</sup>. Prolific enables us to screen potential participants based on their self-reported language backgrounds. To avoid bias, we restricted participation to individuals who learned English as a second language and were not raised in an English-speaking environment, because second language learning is vastly different from first language acquisition. Native speakers may never have experienced conversations like those in the TSCC corpus, and therefore cannot fully understand why a learning point is interesting to a second language learner. For example, if a teacher uses simple terms tailored to the student’s proficiency level, a native speaker annotator might score it as very low in interestingness because they would judge it as too easy from their own perspective.

Upon registering for the study, all participants were provided with detailed guidelines and informed that the data would be anonymized and made available to researchers. Participants were

<sup>3</sup>prolific.com

Level	Count			Average Score		AC2 Agreement		
	Row	Conv	Annotator	INT	EXP INT	INT	EXP INT	INT & EXP INT
Turn	7, 118	65	96	2.10 <sub>0.77</sub>	2.00 <sub>0.71</sub>	0.40 <sub>0.13</sub>	0.39 <sub>0.15</sub>	0.39 <sub>0.13</sub>
Sequence	5, 801	259	48	1.35 <sub>0.80</sub>	1.42 <sub>0.74</sub>	0.58 <sub>0.14</sub>	0.52 <sub>0.15</sub>	0.55 <sub>0.13</sub>

Table 1: The number of turns/sequences, conversations, and annotators. The mean and standard deviation of the average interestingness (INT) and expected interestingness (EXP INT) across the 3 annotators that rate each conversation. The average AC2 agreement of the ratings for the turn/sequence-level annotations of INT and EXP INT. The gray values are the standard deviations.

then directed to our locally hosted annotation platform, which was built using the open-source software *doccano*<sup>4</sup>. Participants were instructed to assume the role of either teacher or student within a given conversation, and to base their ratings on that role, rather than their personal preferences. No participant was assigned both roles within the same conversation.

**Annotator Compensation** A total of 96 participants completed turn-level annotations, while another 48 participants annotated at the sequence-level. Turn-level annotators were compensated at a rate of 7£ per hour. Sequence-level annotators received 8£ per hour, with an additional 3£ bonus for high alignment within their annotation group.

**Enhancing Annotation Quality** We initially conducted a pilot study at the turn-level. As is common in subjective annotation tasks involving engagement or affective judgments, we observed relatively low inter-annotator agreement and high annotator effort, particularly when rating individual messages in isolation. These findings are consistent with prior work highlighting the challenges of obtaining consistent judgments in tasks involving nuanced, interpretive constructs such as interest or curiosity (Rodriguez et al., 2020; Rottger et al., 2022). To increase reliability and mitigate annotator fatigue, we implemented two strategies aligned with the literature in preference learning (Brown et al., 2020).

First, each annotation task included two reference conversations for calibration, selected from examples where previous annotators had shown high agreement. These *reference conversations* served as calibration examples, covering a diverse range of engagement levels and linguistic features (e.g., topic shifts, complexity). The first 11 turns from each reference conversation were prepended and appended to each annotation task.

<sup>4</sup><https://github.com/doccano/doccano>

Second, to reduce subjectivity in scoring, we adopted a comparison-based approach inspired by Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022). Annotators rated each message alongside a “boring” alternative generated by GPT-4o, an automatically rewritten version that preserved content while removing engaging linguistic features (see the example below and in Appendix B). The prompts are shown in Appendix C. This setup allowed annotators to judge interestingness in relative rather than absolute terms, a method shown to reduce cognitive load and improve consistency in preference-based tasks (Clark et al., 2018).

Original Messages

it's break-broke-broken  
it's breaking

Boring Alternative

The correct verb forms for "break" are: break, broke, broken.  
it is currently in the process of breaking

**Sequence-Level Annotations** Despite the aforementioned measures, our pilot studies confirmed that turn-level annotation remained more variable and cognitively demanding than sequence-level annotation. Given that engagement in dialogue often builds across multiple turns, we therefore prioritized sequence-level engagement annotations as they provide a more reliable and informative representation of conversational engagement focused on multi-turn discourse modelling.

To conduct this annotation, we leveraged TSCC V2 labels to segment conversations based on teaching content, ensuring that each sequence reflected a cohesive unit of interaction. As described in Section 3.2, annotators assigned interestingness and expected interestingness scores to each sequence. To enhance annotation consistency and agreement,

also in this case we also incorporated a comparison-based alternative, where annotators evaluated original turns alongside GPT-3.5-generated rewrites to facilitate more consistent engagement judgments.

**Quality Control and Agreement Measures** To maintain annotation quality, we implemented sanity checks and a reward-based incentive system.

First, we conducted a sanity check to identify low-effort annotations by detecting sequences where annotators assigned identical scores to 10 or more consecutive turns were excluded from the dataset. Second, each message was annotated by three annotators, and agreement was measured using Gwet’s AC2 metric (Gwet, 2008, 2014), chosen for its stability in cases of variable annotator agreement. AC2 is computed as:

$$\text{AC2} = 1 - \frac{\sum w_{ij} \cdot f_{ij}}{\sum w_{ij} \cdot e_{ij}} \quad (1)$$

where  $f_{ij}$  is the frequency with which the first rater gives score  $i$  and the second rater gives score  $j$ ,  $e_{ij}$  is the expected frequency under the assumption of statistical independence, and  $w_{ij}$  is the weight for agreement between scores  $i$  and  $j$ . Using an identity matrix as  $W$  would treat both small ( $3 \rightarrow 4$ ) and large ( $1 \rightarrow 4$ ) differences equally as complete disagreement. Thus, we instead applied a linear weight matrix (Eq. 2), which makes disagreement decrease linearly with the distance between categories. It is also necessary to clarify that this choice of weight matrix is part of the evaluation metric and is independent from the data.

$$W_{\text{AC2}} = \begin{bmatrix} 1.00 & 0.75 & 0.5 & 0.25 & 0.00 \\ 0.75 & 1.00 & 0.75 & 0.50 & 0.25 \\ 0.50 & 0.75 & 1.00 & 0.75 & 0.50 \\ 0.25 & 0.50 & 0.75 & 1.00 & 0.75 \\ 0.00 & 0.25 & 0.50 & 0.75 & 1.00 \end{bmatrix} \quad (2)$$

Finally, a reward system was implemented at the sequence-level stage. Annotator groups (of three) achieving an AC2 score of 0.5 or higher on their *interestingness* annotations received a £3 bonus per annotator and were prioritized for future tasks.

Table 1 reports the dataset statistics, average scores, and AC2 agreement of the turn-level and sequence-level annotations. We compute the AC2 of 3 annotators jointly in each annotation task and then average them across all the tasks. Sequence-level annotations have a significantly higher level of agreement than turn-level on INT ( $p = 3.19e^{-8}$ ), EXP INT ( $p = 6.82e^{-5}$ ), and combined scores ( $p = 5.12e^{-7}$ ), indicating the difficulty in evaluating interestingness on a fine-grained level.

### 3.4 Dataset Statistics and Demographics

Turn-level annotations were collected for 64 conversations (25% of the corpus), while sequence-level annotations, focusing solely on teacher messages, were collected for 259 conversations<sup>5</sup>. A total of 5,801 sequences were annotated, with an average of 22.4 sequences per conversation (ranging from 8 to 41). Turn-level annotations consist of 7,118 pairs of scores in total.

Table 2 presents the self-reported age, gender, first language, and English proficiency (measured using the Common European Framework of Reference, CEFR) of participants in the turn- and sequence-level annotation tasks. In both cases, most annotators identified as B2 or C1 level, reflecting an upper-intermediate to advanced command of English. This distribution was intentional, as it aligns with the proficiency levels of students in the original TSCC dataset (i.e., mostly from B1 to C2), ensuring that annotators could reliably adopt the perspective of the original students in the conversations. We also found that conversations with learners of higher proficiency were rated as more interesting, and also that ratings were higher when the annotator’s proficiency matched the proficiency of the learner in the original conversation. Nonetheless, it is possible that the skew towards annotators with relatively high proficiency may introduce some bias, as annotators not actively engaged in language learning might have perceived the teacher’s instruction as overly simplistic. Additionally, the recruitment platform, Prolific, which is primarily used in Europe and North America, naturally favors users with moderate to advanced English proficiency.

## 4 Experiments

In this section, we leverage the new InTrEx dataset to conduct an experimental assessment of engagement in educational dialogue, focusing on two complementary goals: (i) evaluating the extent to which LLMs align with human judgments of interestingness, and (ii) identifying the linguistic features that contribute to perceived engagement.

### 4.1 Reward Modeling with InTrEx

First, we investigate how LLMs, with or without instruction-tuning, assess conversational interestingness. We then fine-tuned Llama3-8B (Dubey

<sup>5</sup>Annotator information can be downloaded from <https://github.com/Xingwei-Tan/InTrEx>

Level	CEFR	Count	Average Age	Gender	First Language (Top 3)
Turn	C2	13	28.08	M (7), F (6)	Greek (4), Polish (3), Portuguese (1)
	C1	36	25.78	M (25), F (11)	Polish (14), Portuguese (8), Italian (5)
	B2	23	30.13	M (10), F (13)	Polish (14), Italian (6), Portuguese (1)
	A2	3	33.33	M (1), F (2)	Polish (1), Greek (1), Spanish (1)
Sequence	C2	9	29.56	M (6), F (3)	Portuguese (5), Greek (2), Polish (1)
	C1	24	27.13	M (12), F (12)	Portuguese (8), Polish (4), Italian (4)
	B2	14	28.21	M (6), F (8)	Polish (7), Estonian (2), Portuguese (2)
	A1	1	24.00	M (1)	Polish (1)

Table 2: The average age and the distributions of gender and first language grouped by the CEFR level of the annotators in the turn-level and sequence-level annotation.

Predictor	AC2
Random (Gaussian)	0.3491
GPT-4	0.4421
GPT-4o	0.4657
Mistral-7B-Base	0.2129
Llama3-8B-Base	0.2584
Mistral-7B-Instruct	0.3608
Llama3-8B-Instruct	0.3646
Mixtral-8×7B-Instruct	0.4549
Mistral-7B-Base (-IntrEx)	0.1587
Llama3-8B-Base (-IntrEx)	0.2340
Mistral-7B-Instruct (-IntrEx)	<b>0.5142</b>
Llama3-8B-Instruct (-IntrEx)	0.5139

Table 3: The average AC2 agreement between the LLMs and the human rating of interestingness. The last four rows are models fine-tuned on IntrEx.

et al., 2024) and Mistral-7B (Jiang et al., 2023) to predict human *interestingness* scores using IntrEx as a multi-class classification task. The model inputs consisted of the conversation history preceding the target message to be scored. The target message was delimited by <target-of-rating> and </target-of-rating> tags. The average of the three annotators’ scores for each message, rounded to the nearest integer, served as the ground truth label. Sequence-level annotations comprised the training set, while turn-level annotations formed the test set.

Table 3 compares the performance of our fine-tuned models against the off-the-shelf LLMs using in-context prompting (prompts are reported in Appendix C). As an additional baseline, we include a random number generator which samples from  $N(2, 1)$  and rounds to the nearest integer. The random baseline has an AC2 agreement of 0.3491 with respect to humans. The frontier proprietary LLMs GPT-4 and GPT-4o achieved AC2 scores of 0.4421 and 0.4657, respectively, outperforming the smaller Mistral-7B and Llama3-8B. However, af-

ter fine-tuning on the IntrEx, Llama3-8B-Instruct and Mistral-7B-Instruct surpass the much larger GPT-4, GPT-4o, and Mixtral (Jiang et al., 2024). This result also shows that the fine-tuned models demonstrate strong generalization capabilities, generating fine-grained, turn-level feedback despite being trained on coarser, sequence-level data.

Notably, base LLMs performed worse than the random baseline, even after fine-tuning (Table 3). This suggests that - since they lack pre-training exposure to explicit rating tasks - base models struggle with instruction following in rating-based contexts. In contrast, instruction-tuned models (e.g., Llama3-8B-Instruct) showed greater alignment with human interestingness scores, emphasizing the importance of instruction tuning for engagement modeling.

## 4.2 Linguistic Predictors of Human Interest

In our second analysis, we examine how linguistic factors, such as concreteness, comprehensibility, and uptake, contribute to perceived engagement in teacher–student interactions. To examine the individual and combined effects of concreteness, comprehensibility, and uptake on interestingness, we computed a range of metrics (Table 4).

- **Concreteness** was measured as the average concreteness rating of the content words in each turn or sequence. The concreteness ratings are from the MRC Psycholinguistic Database (Wilson, 1988). We selected the metric based on the MRC database due to its broader lexical coverage across the conversational vocabulary in our dataset.
- **Gist Inference Score (GIS)** is based on the principle that more comprehensible texts facilitate gist extraction: the ability to form an abstract representation of the content without needing to retain *verbatim* details. Be-

Level	Feature	Interestingness			Expected Interestingness		
		B	SE	<i>t</i>	B	SE	<i>t</i>
Turn	(Intercept)	2.085	0.068	30.718	1.977	0.067	29.733
	<b>Concreteness</b>	-0.039	0.009	<b>-4.478</b>	-0.018	0.009	<b>-2.060</b>
	<b>Coleman-Liau index</b>	0.029	0.011	<b>2.618</b>	0.030	0.011	<b>2.776</b>
	<b>Smog index</b>	0.030	0.010	<b>3.035</b>	0.022	0.010	<b>2.316</b>
	Automated readability index	0.004	0.019	0.201	-0.007	0.019	-0.364
	Spache readability	-0.010	0.016	-0.594	-0.011	0.016	-0.713
	<b>GIS</b>	0.020	0.009	<b>2.384</b>	0.009	0.008	1.115
	<b>GIS<sup>2</sup></b>	-0.028	0.007	<b>-3.996</b>	-0.013	0.007	-1.808
	<b>Lexicon count</b>	0.197	0.010	<b>19.036</b>	0.177	0.010	<b>17.315</b>
	<b>Lexicon count<sup>2</sup></b>	-0.117	0.008	<b>-15.024</b>	-0.093	0.008	<b>-12.051</b>
	<b>LCS</b>	0.028	0.008	<b>3.615</b>	0.009	0.008	1.163
	<b>Student-uptake-teacher</b>	0.020	0.007	<b>2.807</b>	0.015	0.007	<b>2.080</b>
	Cosine similarity	-0.015	0.008	-1.933	-0.013	0.008	-1.684
Sequence	(Intercept)	1.379	0.050	27.818	1.425	0.058	24.586
	Concreteness	0.000	0.012	-0.039	0.018	0.011	1.597
	<b>Flesch reading ease</b>	0.459	0.039	<b>11.891</b>	0.319	0.037	<b>8.694</b>
	<b>Flesch-Kincaid grade</b>	0.677	0.066	<b>10.229</b>	0.539	0.063	<b>8.566</b>
	<b>Coleman-Liau index</b>	0.238	0.025	<b>9.528</b>	0.201	0.024	<b>8.453</b>
	Dale-Chall readability	0.009	0.016	0.566	0.025	0.015	1.691
	Gunning fog	-0.026	0.029	-0.914	-0.017	0.027	-0.612
	<b>Smog index</b>	0.276	0.010	<b>27.897</b>	0.194	0.009	<b>20.617</b>
	<b>Automated readability index</b>	-0.195	0.039	<b>-5.038</b>	-0.254	0.037	<b>-6.907</b>
	<b>Spache readability</b>	-0.093	0.036	<b>-2.622</b>	-0.036	0.034	-1.073
	<b>Student-uptake-teacher</b>	-0.055	0.009	<b>-6.254</b>	-0.035	0.008	<b>-4.186</b>
	<b>propTinS</b>	0.113	0.010	<b>11.291</b>	0.075	0.009	<b>7.934</b>
	<b>Cosine similarity</b>	-0.046	0.011	<b>-4.219</b>	-0.088	0.010	<b>-8.554</b>

Table 4: Combined model predicting turn-level or sequence-level ratings respectively. All predictors are scaled, so coefficients (**B**) are standardised. **SE** is the standard error of the coefficient. *t* is the t-statistic. Significant predictors are highlighted in bold (i.e.,  $|t| \geq 2$ ).

- cause the GIS score formula inherently penalizes concreteness, and given their moderate negative correlation, we analyzed concreteness separately as well as in conjunction with the GIS score. Critically, the GIS score has been empirically validated, at least for medical texts, with evidence suggesting that higher GIS scores correlate with greater human comprehension (Wolfe et al., 2019; Dandignac and Wolfe, 2020).
- **Flesch reading ease** indicates how easy the text is to read. The lower the score, the tougher it is to read. The formula takes into account the total number of words, sentences, and syllables.
  - **Flesch-Kincaid grade** is similar to the Flesch Reading Ease, using the same variables in the formula, but it is more extensively used in education. A higher score indicates the text is suitable for a higher grade, thus more difficult.
  - **Coleman-Liau index** depends on the average number of letters per 100 words and the average number of sentences per 100 words.
  - **Dale-Chall readability** uses a list of 3000 words that groups of fourth-grade American

students rated on a difficulty scale.

- **Gunning fog index** determines the years of formal education required by the reader to understand the text. It is based on the number of words, sentences and complex words.
- **Smog index** uses the number of sentences and polysyllables to determine the years of education needed to understand text.
- **Automated readability index** determines the comprehensibility of a text based on the number of characters, words and sentences.
- **Spache readability** rates the comprehensibility of text based on the number of common everyday words it contains.
- **Longest Common Subsequence (LCS)** identifies common words in consecutive student and teacher turn.
- **Student-uptake-teacher** assess the extent to which the next speaker built upon the previous speaker’s contribution for two consecutive turns. This metric was derived from a BERT model (Wang and Demszky, 2024a) trained on a dataset of educational conversations (Demszky et al., 2023). The score represents the student’s uptake of the teacher’s turn.



- **propTinS** is the proportion of words the teacher repeated from the student’s message.
- **Cosine similarity** is the similarity between *mxhai-embed-large-v1* (Lee et al., 2024) embeddings of consecutive teacher/student turns.

**Concreteness.** We found that concreteness was negatively correlated with interestingness, which is in line with emerging evidence that overly simple content may reduce engagement (Oudeyer et al., 2016; Dubey and Griffiths, 2020). This finding is consistent across turn- and sequence-level.

**Comprehensibility.** Comprehensibility was assessed via a wide-range of metrics (i.e., GIS, Flesch reading ease, Flesch-Kincaid grade, Coleman-Liau index, Dale-Chall readability, Gunning fog index, Smog index, Automated readability index, and Spache Readability) We first run comparisons between a simple linear regression model and a regression model including both a linear and a quadratic predictor (linear and quadratic predictors were always orthogonal polynomials) for each comprehensibility feature separately. We then included all significant (i.e.,  $|t| \geq 2$ ) predictors in a single linear mixed-effect model with random (intercept) effects for conversation and annotation project (i.e., subset of 3 annotators who completed the same annotation task). This was done twice: with either interestingness or expected interestingness as the outcome variable, and we only retained those metrics that were robust predictors in both models.

At both turn- and sequence-level, several readability metrics were positively related to interest, suggesting that conversations that are more suitable for a higher grade tended to be more interesting for our annotators. Note that, at the sequence-level, some readability metrics had instead a negative relation with interest, suggesting the relation between comprehensibility and interest is complex. In addition, but only at the turn-level, we observed an inverted U-shaped relation with turn length, i.e., lexicon count, the GIS score, and gunning fog, a traditional readability metric. Such an inverted U-shaped relation is predicted by several computational and psychological theories of interest (Oudeyer et al., 2016; Dubey and Griffiths, 2020), and suggests that both very simple and very complex messages lower interest levels, compared to messages that challenge the reader “just enough”.

**Uptake.** While concreteness and comprehensibility are widely applicable to different types of text, the third feature we investigated, *uptake*, is spe-

cific to conversational turns. Qualitative research using conversation analysis has long emphasized the importance of uptake, particularly in educational settings (Huth, 2011; Walsh, 2013). The degree to which a teacher reuses elements of a student’s turn provides reassurance that the student is being heard and encourages further interaction, thereby enhancing the learning process and boosting student confidence. See et al. (2019) explored whether controlling for response-relatedness (measured via embedding similarity) improves human judgments of *interestingness* in human-LSTM network conversations. Inspired by this work, we also investigated response-relatedness, calculating several uptake measures: LCS, propTinS, model-based estimates of conversational uptake, and embedding-based cosine similarity between turns.

In both turn-level and sequence-level analyses, we found positive correlations between some uptake measures and interest (*LCS* and *student-uptake-teacher* in turn-level analyses, *propTinS* in sequence-level analyses). However, cosine similarity was a negative predictor in both turn-level and sequence-level analyses, and student-uptake-teacher was a negative predictor in sequence-level analyses. This seemingly contradictory finding could suggest that introducing novel ideas, which lowers similarity between successive turns or sequences, may also be important to promote interest, and that uptake of the teacher by the student may index a different phenomenon when it is computed at the turn-level vs. the sequence-level.

## 5 Conclusion

We introduced InTrEx, the first dataset that annotates both interestingness and expected interestingness for teacher–student dialogues, extending TSCC with sequence-level labels to track how engagement evolves across a lesson. We collected ratings from more than a hundred L2 learners and designed a comparison-based annotation, achieving substantially higher reliability at the sequence-level than the turn-level. Small, instruction-tuned LLMs fine-tuned on InTrEx surpassed GPT-4/4o in predicting human interest ratings, demonstrating the relevance and impact of high-quality labels for the task. Future work will be able to leverage InTrEx for training reward models for generation, and extend the approach to new learner groups and domains.

## Limitations

**Annotator Proficiency Bias** While our annotators were second-language English speakers, most had at least B2 proficiency. This choice was intentional to align with the proficiency level of the original students in the TSCC dataset and to ensure annotators could reliably interpret the conversations and follow the task instructions. However, this may underrepresent the perspective of lower-proficiency learners, whose criteria for engagement might differ. Capturing those perspectives would require task simplification and dedicated data collection, which we consider an important direction for future work.

**Subjectivity of Interestingness** Interestingness is inherently subjective, and inter-annotator agreement is limited by individual variation in interests and background knowledge. We mitigated this through (i) targeted recruitment of language learners with sufficient comprehension skills, (ii) consistent task instructions emphasizing pedagogical engagement rather than personal preference, and (iii) a comparison-based annotation framework that anchors judgments against low-interest alternatives. Nonetheless, subjectivity cannot, and probably should not, be eliminated entirely, and our results should be interpreted accordingly.

**Generalizability to Other Domains** The InTrEx dataset focuses exclusively on teacher–student interactions within English-as-a-second-language learning contexts. As such, the findings may not fully generalize to informal peer conversations, other educational domains (e.g., math), or multilingual scenarios. While our controlled setting allowed us to isolate linguistic drivers of engagement, future work could extend this approach to broader conversational contexts, including long-term dialogue or cross-domain applications.

**Model Evaluation Scope** Our evaluation of LLMs is limited to predicting interest ratings on human-authored conversations. While this serves as a proxy for alignment with human judgments, we do not assess whether fine-tuned models can generate more engaging conversations themselves. Exploring generation quality and incorporating direct preference optimization (e.g., via DPO) would require new rounds of human evaluation and fall outside the scope of this initial dataset release.

## Ethics Statement

Prior to recruiting participants, we obtained ethics approval from our institutional review board, with the understanding that all annotations would be anonymized before public release. While demographic data were collected, these data are not linked to individual participants. A clear withdrawal procedure was provided to all participants. Annotators were compensated fairly for their work, commensurate with the task’s complexity.

## Acknowledgements

Xingwei was supported by the Warwick Chancellor’s International Scholarship while conducting this work. Chiara Gambi was supported by a Leverhulme Trust Research Project grant (RPG-2023-067). Gabriele Pergola was partially supported by the ESRC-funded project *Digitising Identity: Navigating the Digital Immigration System and Migrant Experiences*, as part of the Digital Good Network. This work was conducted on the Sulis Tier-2 HPC platform hosted by the Scientific Computing Research Technology Platform at the University of Warwick. Sulis is funded by EPSRC Grant EP/T022108/1 and the HPC Midlands+ consortium.

We thank the anonymous reviewers for their helpful comments. The writing of this paper received minor language-polishing suggestions from Gemini. In addition, parts of our experimental code were drafted or refactored with assistance from GitHub Copilot; all final implementations were manually reviewed and verified by the authors.

## References

- Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay, and Alfredo Urzua. 2004. *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. Test of English as a Foreign Language.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

- 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.
- Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. [The teacher-student chatroom corpus](#). In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.
- Ronald Carter and Michael McCarthy. 1997. *Exploring spoken english*, volume 2. Cambridge University Press.
- Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? using the “elochoice” package to assess pairwise comparisons of perceived physical strength. *PLoS one*, 13(1):e0190393.
- Mitchell Dandignac and Christopher R Wolfe. 2020. Gist inference scores predict gist memory for authentic patient education cancer texts. *Patient Education and Counseling*, 103(8):1562–1567.
- Dorottya Demszky, Jing Liu, Heather C Hill, Shyamoli Sanghi, and Ariel Chung. 2023. Improving teachers’ questioning quality through automated feedback: A mixed-methods randomized controlled trial in brick-and-mortar classrooms. edworkingpaper no. 23-875. *Annenberg Institute for School Reform at Brown University*.
- Ed Donnellan, Sumeyye Aslan, Greta M Fastrich, and Kou Murayama. 2022. How are curiosity and interest different? naïve bayes classification of people’s beliefs. *Educational Psychology Review*, 34(1):73–105.
- Zoltán Dörnyei and Ema Ushioda. 2021. *Teaching and researching motivation*. Routledge.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Rachit Dubey and Thomas L Griffiths. 2020. Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, 127(3):455.
- JA Goris, EJP Denessen, and LTW Verhoeven. 2019. Effects of content and language integrated learning in europe a systematic review of longitudinal experimental studies. *European educational research journal*, 18(6):675–698.
- Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology*, 61:29–48.
- Kilem L. Gwet. 2014. [Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters](#).
- Pedram Hosseini, Christopher Wolfe, Mona Diab, and David Broniatowski. 2022. [GisPy: A tool for measuring gist inference score in text](#). In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 38–46, Seattle, United States. Association for Computational Linguistics.
- Thorsten Huth. 2011. Conversation analysis and language classroom discourse. *Language and Linguistics Compass*, 5(5):297–309.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. [Mistral of experts](#).
- Celeste Kidd and Benjamin Y Hayden. 2015. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.
- Andreas Krapp. 1994. Interest and curiosity. the role of interest in a theory of exploratory action. In *Curiosity and exploration*, pages 79–100. Springer.
- Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada.
- Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. [Open source strikes bread - new fluffy embeddings model](#).
- Sunjung Lee and Diana Pulido. 2017. The impact of topic interest, l2 proficiency, and gender on efl incidental vocabulary acquisition through reading. *Language Teaching Research*, 21(1):118–135.
- George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75.

- Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The spoken bnc2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.
- Sarah M. Lupo, Laura S. Tortorelli, Marcia Invernizzi, Ji Hoon Ryoo, and John Z. Strong. 2019. [An exploration of text difficulty and knowledge support on adolescents’ comprehension](#). *Reading Research Quarterly*.
- Chen Lyu, , and Gabriele Pergola. 2024. [SciGisPy: a novel metric for biomedical text simplification via gist inference score](#). In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 95–106, Miami, Florida, USA. Association for Computational Linguistics.
- A-M Masgoret and Robert C Gardner. 2003. Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by gardner and associates. *Language learning*, 53(S1):167–210.
- Kou Murayama. 2022. A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. *Psychological Review*, 129(1):175.
- Dang Nguyen, Jiuhai Chen, and Tianyi Zhou. 2024. [Multi-objective linguistic control of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4336–4347, Bangkok, Thailand. Association for Computational Linguistics.
- Anne O’Keeffe and Steve Walsh. 2012. Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory*, 8(1):159–181.
- OpenAI, 2024. 2024. [Gpt-4 technical report](#). OpenAI.
- P-Y Oudeyer, Jacqueline Gottlieb, and Manuel Lopes. 2016. Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. *Progress in brain research*, 229:257–284.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Christopher Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Kwabena Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.
- Gabriele Pergola, Lin Gui, and Yulan He. 2019. Tdam: A Topic-Dependent Attention Model for Sentiment Analysis. *Information Processing & Management*, 56(6):102084.
- Gabriele Pergola, Lin Gui, and Yulan He. 2021. [A disentangled adversarial neural topic model for separating opinions from plots in user reviews](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2870–2883, Online. Association for Computational Linguistics.
- Emily Grossnickle Peterson and Suzanne Hidi. 2019. Curiosity and interest: current perspectives. *Educational Psychology Review*, 31(4):781–788.
- Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. 2020. [Information seeking in the spirit of learning: A dataset for conversational curiosity](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, Online. Association for Computational Linguistics.
- Carolyn P. Rosé, Diane Litman, Dumisizwe Bhembe, Kate Forbes, Scott Silliman, Ramesh Srivastava, and Kurt VanLehn. 2003. [A comparison of tutor and student behavior in speech versus text based tutoring](#). In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 30–37.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Mark Sadoski. 2001. Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review*, 13:263–281.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.
- Rita C Simpson, Sarah L Briggs, Janine Ovens, and John M Swales. 2002. The michigan corpus of academic spoken english. *Ann Arbor, MI: The Regents of the University of Michigan*.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.
- Xingwei Tan, Chen Lyu, Hafiz Muhammad Umer, Sahrish Khan, Mahathi Parvatham, Lois Arthurs, Simon Cullen, Shelley Wilson, Arshad Jhumka, and Gabriele Pergola. 2025a. [SafeSpeech: A comprehensive and interactive tool for analysing sexist and abusive language in conversations](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System*

- Demonstrations*), Albuquerque, New Mexico. Association for Computational Linguistics.
- Xingwei Tan, Yuxiang Zhou, Gabriele Pergola, and Yulan He. 2025b. [Cascading large language models for salient event graph generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2223–2245, Albuquerque, New Mexico. Association for Computational Linguistics.
- Duolingo Team. 2023. [Introducing duolingo max, a learning experience powered by gpt-4](#).
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. [Lamda: Language models for dialog applications](#).
- Tan Bee Tin. 2009. Features of the most interesting and the least interesting postgraduate second language acquisition lectures offered by three lecturers. *Language and Education*, 23(2):117–135.
- Etsuko Toyoda and Richard Harrison. 2002. Categorization of text chat communication between learners and native speakers of Japanese.
- Steve Walsh. 2013. *Classroom discourse and teacher development*. Edinburgh University Press.
- Rose Wang and Dorottya Demszky. 2024a. [EduConvoKit: An open-source library for education conversation data](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 61–69, Mexico City, Mexico. Association for Computational Linguistics.
- Rose E Wang and Dorottya Demszky. 2024b. [Educonvokit: An open-source library for education conversation data](#). *arXiv preprint arXiv:2402.05111*.
- William P Wharton. 1988. Imagery and the comprehension of college history texts: Free response measure. *Imagination, Cognition and Personality*, 7(4):323–333.
- Margaret Wilson. 1988. [The mrc psycholinguistic database: Machine readable dictionary, version 2](#). *Behavioural Research Methods, Instruments, & Computers*, 20(1):6–11.
- Christopher R Wolfe, Mitchell Dandignac, and Valerie F Reyna. 2019. A theoretically motivated method for automatically evaluating texts for gist inferences. *Behavior research methods*, 51:2419–2437.

## A Fine-tuning LLMs on Human Interest Annotations

We tested several settings of training models to predict the interestingness scores. Our first attempt is to train the models to predict the scores as a regression task, but the models often ignore the range of the scores, producing scores lower than 0 or greater than 4. We also tried adding instructions to the input to describe that “Please rate the level of interestingness of the given message based on the context of the whole conversation.”, but the resulting models are inferior than the ones trained without this instruction. We compared the performance of training with the base model versus the instruction-tuned version of Llama3 and Mistral, and found the instruction-tuned models have higher agreement. The Mistral and Llama3 were trained for 3 epochs with a learning rate of  $5e - 6$ .

## B Boring Alternative

### Origin Conversation

TEACHER: OK good...this is what means to be addicted i.e. you can't stop ...good! OK so a bit more about how/why this is bad.....!  
STUDENT: as a consequence, young people don't have time to do some useful things for their life, such as being social and reading.  
TEACHER: OK great .... a clear reason about why this is bad ...hold on a sec....

### Boring Alternative

TEACHER: Please elaborate further on the negative implications of this behavior.  
STUDENT: as a consequence, young people don't have time to do some useful things for their life, such as being social and reading.  
TEACHER: Please wait momentarily while I process the information...

## C Prompts

### Prompt for generating boring alternatives

Instruction: given a text message from a teaching session between a teacher and a student, please provide a more straightforward and less engaging version. Strip away any colourful language or additional context to make the message as boring as possible. Please keep the main information from the message.

The prompt used for instructing off-the-shelf instruction-tuned LLMs to rate the turn-level INT and EXP INT:

### Prompt for instructing instruction LLMs

Can I please ask you to rate a student-teacher conversation based on how interesting they feel? Provide a rating between 0 to 4, with 0 being boring and 4 being most interesting. This conversation involves a student learning English as a second language from a teacher. Assume the role of the student. Rate as follows: INT: 'interest in the teacher's reply,' EXP\_INT: 'expected interest in the next conversation,' REASON: 'justify the rating.' Consider previous conversations. Next Dialogue: [Teacher and student dialogue snippet]

The prompt used for instructing off-the-shelf base LLMs to rate the turn-level INT and EXP INT:

### Prompt for instructing base LLMs

The student said [Teacher and student dialogue snippet]. To evaluate how interesting the message of the student is, the teacher gave an interestingness score from: 0 = not interesting, 1 = slightly interesting, 2 = interesting, 3 = very interesting, 4 = extremely interesting. The teacher gave:

## D Metric Implementation

Traditional readability metrics were computed using the *textstat* library<sup>6</sup>). The GIS score was computed from the GisPy package (Hosseini et al., 2022), a Python toolkit for extracting psycholinguistic features. The model-based estimates of conversational uptake are based on Edu-Convokit (Wang and Demszky, 2024b).


## E Dataset Information

The dataset is intended for research purposes only and is released in compliance with the original access conditions set by Prolific and the ethics approval guidelines. The dataset has been fully anonymized, with all personally identifiable information removed. Participants were informed about anonymization procedures and data protection measures. The dataset is documented with details on domains, languages, linguistic phenomena, and demographic information, ensuring transparency and reproducibility.

## F Annotation Interface

In this section, we provide more details of our recruitment and annotation interface; see Figure 5. Figure 3 and 4 show the recruitment page on Prolific.

<sup>6</sup><https://pypi.org/project/textstat/>



## Evaluate Interestingness in English Learning Conversations

By warwick.ac.uk

£8.00 • £8.00/hr
🕒 1 hour
👤 3 places

**You are invited because your annotations in our previous tasks are high-quality. You could go straight to the annotation platform if you still remember your password. Otherwise, please message the researcher.** In this study, you will be asked to annotate a series of online text conversations between a teacher and a student. In each conversation, you will be asked to evaluate the teacher's messages. For example, you should imagine being the student who took part in this conversation, put yourself in the student's shoes and annotate how interesting you find the messages produced by the teacher.

For each message, you will need to give three labels:

1. **an interestingness score for the current messages you are reading;**
2. **an expectation of the interestingness score for the future messages from the teacher in the conversation.**
3. **choose whether the alternative responses are more or less interesting than the original responses**

The interestingness score represents how much the message catches your attention or arouses your curiosity and interest. The expectation of interestingness score represents your anticipation of how interesting you will find the next messages (on the next page; please provide this score **before** you read the messages on the next page).

**There is a potential bonus of 3£ (max payment 11£/hour)!**

In addition to the standard payment for your annotations, **we offer rewards for high-quality annotations.** The quality of the annotations will be assessed through a series of sanity checks to avoid irrelevant or inaccurate annotations and ensure the average alignment with a large pool of participant submissions.


Should the agreement level within a group exceed our standard threshold, each member will receive **a bonus of 3£**, in addition to the standard payment of £8. Furthermore, participants who produce high-quality work will be invited to undertake more annotation tasks. The same rewards structure will be in place for these subsequent rounds of annotation.

Devices you can use to take this study:

Desktop
  Mobile
  Tablet

[Open study link in a new window](#)

Figure 3: The recruitment page of sequence-level annotation on Prolific.



## Evaluate Interestingness in English Learning Conversations

By warwick.ac.uk

£7.00 • £7.00/hr
🕒 1 hour
👤 7 places

In this study, you will be asked to annotate a series of online text conversations between a teacher and a student. In each conversation, you will be assigned the viewpoint of either the student or the teacher. For example, if you are assigned the viewpoint of the student, you should imagine to be the student who took part in this conversation, put yourself in the student's shoes and annotate how interesting you find the messages produced by the teacher. And vice versa, if you are assigned the viewpoint of the teacher, you should imagine being the teacher and annotate how interesting you find the messages produced by the students.

For each message, you will need to give three labels:

1. **an interestingness score for the current message of the target;**
2. **an expectation of the interestingness score for the next message produced by that target in the conversation;**
3. **choose whether the alternative response is more interesting than the original response.**

The interestingness score represents how much the message catches your attention or arouses your curiosity and interest. The expectation of interestingness score represents your anticipation of how interesting you will find the next message produced by that person (before you have read it).

Devices you can use to take this study:

Desktop
  Mobile
  Tablet

[Open study link in a new window](#)

Figure 4: The recruitment page of turn-level annotation on Prolific.

TEACHER: I've finished = good I finished = not wrong but much much less common  
 STUDENT: ok  
 TEACHER: I'm finished = also OK but again, less common  
 TEACHER: I thought that might be your next question- I was trying to be clever and pre-empt it (it's a common one!)  
 STUDENT: pre-empt??  
 STUDENT: predict

==== Is the following alternative more interesting? ====

TEACHER: I've finished = acceptable. I finished = less common but not incorrect  
 STUDENT: ok  
 TEACHER: I'm finished = acceptable, though less frequently used  
 TEACHER: I anticipated that could be your subsequent inquiry. I attempted to anticipate it cleverly (it's a frequently asked one!).  
 STUDENT: pre-empt??  
 STUDENT: predict

**Progress**

Total 103  
 Complete 0  
 0%

Key	Value
conversation_id	044

Figure 5: The user interface of a page in the sequence-level annotation task.