Evaluating LLM-generated Explanatory Dialogue Turns through Dialogue Completion

Anonymous ACL submission

Abstract

Human dialogues frequently feature explanations when it comes to conveying ideas and engaging in discourses. Synthetic explanatory dialogues offer potential for various applications such as dialogue systems and model self-rationalization. However, synthetic dialogues are typically regarded as inferior in quality compared to human ones. We investigate large language models' capability of completing a missing dialogue turn within a given context of an explanatory conversation. We conduct experiments over three datasets, which cover both natural and synthetic explanatory dialogues, and apply two test suites for evaluation. While the evaluation confirms the quality gap between human and synthetic dialogues, LLM-generated turns are found to outperform human ones in fluency and grammatical accuracy. Moreover, while each of the three investigated models demonstrates distinct strengths and weaknesses on the task, their performance can be consistently improved through prompt-based refinement methods.

1 Introduction

In pursuit of interpreting model behavior under the notion of explainable artificial intelligence (XAI), explanation not only plays opportunities for rationalizing models' decision-making (Lakkaraju et al., 2022; Feldhus et al., 2023). More generally, explanation plays an important role in conceptualizing ideas (Miller, 2017). The enhanced natural language generation (NLG) capability of large language models (LLMs) enables them to bring forward more interactive conversation with human users and even perform explanatory dialogues. Nevertheless, dialogue is by nature complex in terms of back-and-forth exchanges and contextual information regarding speakers or space. The dynamic flow of utterances poses challenges in terms of capturing features and aspects to be modeled. Aspects such as providing statements

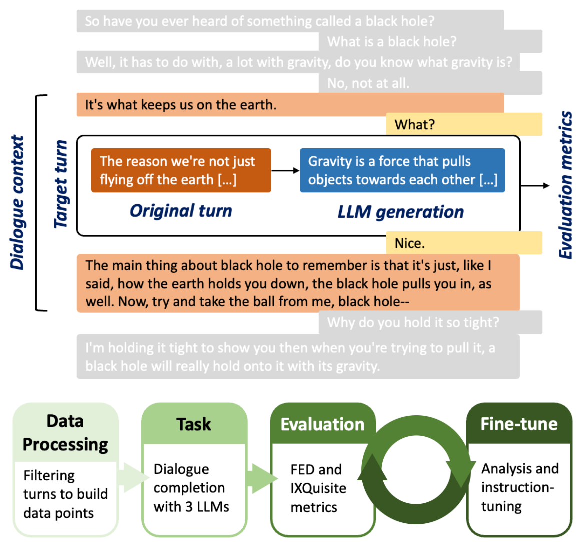


Figure 1: The summary of the dialogue completion task (top) and the pipeline of the experiments (bottom).

and interacting with an interlocutor are difficult for conventional conversational agents, while the more advanced approaches continue to suffer from these challenges very often (Ni et al., 2023).

Focusing on explanatory contexts, our study explores the dialogue completion task (Fig. 1) and deals with how models and human behave differently in a dialogue, as well as the measurement of such differences (Dai et al., 2022; Li et al., 2023; Wang et al., 2024). In light of these motifs, we investigate LLMs' capability of performing explanation on given topics in conversational scenarios, contributing the following: **(1)** We explore how different prompting strategies alters the output space of the dialogue completion task. **(2)** We compare synthetic vs. human and colloquial vs. written dialogues turns using two test suites containing a range of reference-free metrics evaluating dialogue quality, FED (Mehri and Eskénazi, 2020a) and IXQUISITE (Feldhus et al., 2024) (§3.3). **(3)** We make analyses on performance of dialogue completion task, discovering that LLMs tend to phrase information fluently and yet struggle to engage effectively with the interlocutor when it comes to providing explanations in a dialogue form. **(4)** We instruct LLMs to refine their own outputs in pursuit of resembling human explanatory dialogues, thereby confirming LLMs' capability of self-refinement (Madaan et al., 2023; Zhang et al., 2023) (§3.4) given additional information regarding the score gap between task output and the corresponding original turns.[1]

2 Background

2.1 Explanatory dialogue

Under conversation scenarios, explanatory dialogues (Figure 6 in the Appendix) typically exhibit linguistic features distinct from plain text at syntactic, semantic, and pragmatic level to owing to the interaction between speakers. Alshomary et al. (2024) hypothesized a successful explanation to be based on both explanation moves and dialogue acts; that is, effective explanations rely on not only elaborating the topic but the corresponding pragmatic formulation in the back-and-forth interaction.

Real-life scenario Targeting mutual understanding, explanatory dialogues demonstrate high diversity for the purpose of adapting to different target audience. The educational scenario (Liu et al., 2024) is one of the contexts in which explanations most frequently occur (Demszky and Hill, 2023; Kwon et al., 2024): In such a setting, the expertise level and familiarity with a topic becomes an essential concern in studying such conversations (Wachsmuth and Alshomary, 2022).

2.2 Synthetic dialogues

Previous studies reported a gap between human and synthetic dialogues (Dai et al., 2022; Li et al., 2023; Stacey et al., 2024) in that human conversations were commonly considered more commonly applicable. Although recent studies indicated that model-generated explanatory dialogues could be more favorable than human ones, LLMs were regarded not as replacement of human experts but instead an augmentation of experts' explanation capabilities (Li et al., 2024). Similarly, Dai et al. (2022) suggested that, while synthetic data couldn't thoroughly replace human dialogue data, model-generated dialogues may already benefit interactive conversation systems in fine-tuning models towards producing more natural dialogues.

Tack and Piech (2022) applied LLMs to simulate real-world teaching scenarios. Through treating LLM as a teacher agent interacting with a student, their chatbot could deliver decent teaching quality and yet was outperformed by human teachers in interacting with students according to human evaluation. Liu et al. (2024) focused on student behaviors and found LLMs capable of following instructions to simulate students. Concerning five personality traits, LLMs could diversify their output in a personality-aware simulation.

2.3 Evaluating synthetic dialogues

Evaluating dialogue is considered challenging owing to the interactive nature involving speakers and turns. Corresponding criteria cover coherence, participation, and engagement (Adiwardana et al., 2020). Conventional methods for evaluating NLG output are therefore insufficient for capturing the quality of conversational flow as how human would perceive it (Deriu et al., 2021). The failure of conventional reference-dependent metrics such as F_1 and BLEU on evaluating dialogues directed prior studies to develop more advanced approaches that remain robust under the dynamic conversational expressions (Zhang et al., 2021; Ma et al., 2022; Mendonça et al., 2024). USR (Mehri and Eskénazi, 2020b) and FED (Mehri and Eskénazi, 2020a) performed automatic evaluation for turn-level and dialogue-level aspects based on `DialoGPT` without the need for a reference or ground truth. From an alternative perspective, Feldhus et al. (2024) proposed IXQUISITE, a didactics-inspired suite of metrics which targets explanatory dialogues and employs count-based methods to track linguistic features.

2.4 Rationalization and self-refinement

Following early work of free-text rationalization (Camburu et al., 2018; Rajani et al., 2019), more complex prompting methods were proposed in recent years in pursuit of rationalizing models' decision-making. Chain-of-thought prompting (Wei et al., 2022) employed a multi-step reasoning process, while Madaan et al. (2023) introduced self-refine prompting, in which LLMs are recursively prompted for and with feedback concerning their own output, yielding more favorable responses for both human readers and according to automatic metrics.

[1]The code is available at https://anonymous.4open.science/r/dialog_completion-FBF9

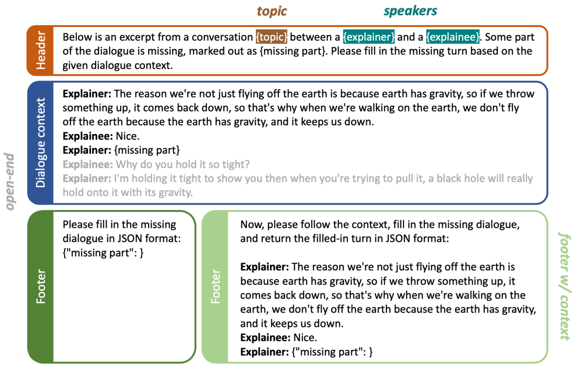


Figure 2: Prompt structure for the explanation dialogue completion task with all the tested prompt variables marked out.

3 Experiments

We instruct LLMs to complete explanatory dialogues. Figure 1 (bottom) presents an overview of our workflow, which is composed of four major stages: Data processing (§A.1), Dialogue completion task (§3.1 & §3.2), Evaluation (§3.3), and Fine-tuning (§3.4). The experiments cover three English-language datasets featuring two-agent explanatory dialogues: ReWIRED (Feldhus et al., 2024), WikiDialog (Dai et al., 2022), and ELI5-dialogues (Alshomary et al., 2024). Details and sample data points are provided in Appendix A.1.

3.1 Task

The dialogue completion task requires LLMs to fill in a removed target turn uttered by the explainer from a segment of a given dialogue context. Formally, an entire dialogue (D) initially consists of n turns (T): $D = T_1, T_2, ..., T_n$.

We select the subset of turns that are longer than minimum length l: $D(l) = i_1, i_2, ..., i_m$.

Finally, for each target turn T_i, a window of length w is applied to retrieve its surrounding dialogue context. LLMs are instructed to fill in the removed target explainer turn t_i in the segmented dialogue context $d_i(w)$, which consists of t_i and its surrounding turns: $(T_{i-w}, ..., T_{i-1}, t_i, T_{i+1}, ..., T_{i+w})$. With LLM-generated turn t'_i replacing the removed t_i turn, the dialogue context is completed as $d_i(w)' = (T_{i-w}, ..., T_{i-1}, t'_i, T_{i+1}, ..., T_{i+w})$.

Figure 9 in the Appendix visualizes the above process and further demonstrates how data instances are built out of the original, unprocessed (*raw*) dialogue: Suitable turns are first selected and then joined by their respective surrounding context retrieved from the raw dialogue. In the prompt, the target turn is replaced with a placeholder which LLMs are instructed to fill in.

Practically, we control the data instance building process with two variables: length (l) and window (w). l defines the minimum number of tokens for a turn to be selected; taking into account that LLM-generated texts tend to fall into a certain length, the variable is capable of filtering out short turns such as *Could you say that again?* or *You're right, I could totally agree*, which can naturally occur in conversations between human. w determines the number of turns prior to and following the target turn in the dialogue context. Since all datasets feature only two speaking agents, we set w to 2 throughout the experiments.[2] Following a rough analysis on the output from a trial run (Appendix B), l is set to 30 tokens.

3.2 Prompting

We conduct experiments with `Mistral-0.3 7B` (Jiang et al., 2023), `Llama-3.1 8B` (Grattafiori et al., 2024), running on NVIDIA A100 GPU. `Claude-3 Haiku` (Anthropic, 2024) (Version `20240307`) is accessed through Anthropic API. All models are of the smallest size within their respective model family. The prompt design is shown in Fig. 2 with an example dialogue from WIRED. Three prompt variables are explored:[3]

Topic. The variable determines whether to establish the dialogue topic in the task description. The dialogue topics are only provided in WIRED / ReWIRED and WikiDialog.

Speakers. By default, the dialogue agents are called *explainer and explainee* in the prompt. The *speakers* variable aims to specify speakers in the task description, referring mainly to the five levels of expertise covered in the WIRED / ReWIRED dataset. For the WikiDialog dataset, applying this variable would lead to mentioning *explainer model and explainee model*, while no changes are applied to the ELI5-dialogues dataset.

Open-end. This variable is used for remove the turns occurring after the target turn and could thereby reduce the context considerably.

WikiDialog and ELI5-dialogues are subsampled to approximate the size of the ReWIRED

[2] Setting w to an odd number would result in the dialogue context starting and ending with an explainee turn. A higher w such as 4, however, could lead to an overly lengthy context.

[3] App. B documents our trial run with a fourth setting.

dataset, which contains 85 dialogues. In the original WIRED corpus, a dialogue has 23.8 turns on average, while most dialogues include fewer than 10 turns in the other two datasets. Therefore, 255 dialogues are randomly selected from each of WikiDialog and ELI5, creating test splits that are three times the size of ReWIRED to ensure a comparable number of target turns for the task.

3.3 Evaluation

In order to evaluate the dialogues, FED (Mehri and Eskénazi, 2020a) and IXQUISITE (Feldhus et al., 2024) (§2.3) are applied to measure the differences between the model-generated turns and their corresponding original ones. On every data point, the two test suites rate the original dialogue and the model-completed dialogue separately, deriving two scores for 24 aspects in total.

FED assesses explanatory dialogues with multiple positive and negative feedback utterances that could reflect the perceived quality for 18 aspects[4]: For example, *Cool! That sounds super interesting* and *That's really boring* are respectively considered positive and negative ones for the "interesting" aspect, which features six such utterances. While around half of the aspects include both positive and negative utterances, the others feature only the negative ones. The score of each aspect is determined by the likelihood for `DialoGPT`, a pre-trained dialogue response generation model, to respond with these utterances to the turn being evaluated.

IXQUISITE (Feldhus et al., 2024) covers seven annotation-dependent acts-related aspects and seven reference-free numerical ones. Unlike metrics such as USR and FED, which rely on pre-trained embedding, IXQUISITE employs numerical methods to track linguistic features including lexical complexity, synonym density, coherence.

3.4 Instruction-tuning

Since LLMs were shown to imitate human personality through prompt-based methods (Liu et al., 2024), we examine to what extent can LLMs be instructed to shorten the gap between model output and natural utterances (Jia et al., 2024; Madaan et al., 2023). For that, we use the quality measurements from §3.3 as recursive feedback in the form of scores. Since most aspects covered by FED and IXQUISITE do not share the same scale, the scores are first normalized and then subtracted to measure the difference for the individual feature of a given data point: $\Delta f = \frac{f_m - \mu_F}{\sigma_F} - \frac{f_o - \mu_F}{\sigma_F}$, where f_m and f_o respectively represent the feature score of model output and the original dialogue. In the standard score normalization, mean value μ_F and standard deviation σ_F are calculated from all scores for the given feature $\Sigma \frac{F_m + F_o}{n_{(m+o)}}$, i.e., including both scores of model output and original turns.

For each instance, we then retrieve n worst-performing features with the lowest Δf. We set n to 3, so that many features are included without building an overly lengthy instruction. These aspects are then mapped to descriptions[5] to form instructions. In addition to the rule-based instructions, two **strategies** are tested for the overall prompt structure (Fig. 18 in Appendix): **(1)** Modifying the original prompt (Fig. 2) by adding the instruction to the end of the header; **(2)** Rewriting the header such that the model has to revise the filled-in turn instead, inspired by prior work in LLM-based free-text evaluation (Chiang and Lee, 2023; Xu et al., 2023; Jiang et al., 2024).

	ReWIRED	WikiDialog	ELI5-dialogues
Built data points (l=30, w=2)	743	313	583
`Mistral-0.3`	398 (53.6 %)	214 (68.3 %)	318 (54.5 %)
`Llama-3.1`	521 (70.1 %)	272 (86.9 %)	448 (76.9 %)
`Claude-3`	524 (70.5 %)	246 (78.6 %)	234 (40.1 %)

Table 1: Data points and percentage of task accomplishment across models and datasets. The numbers are averaged from the four prompted variables (App. D).

4 Results

4.1 Dialogue completion

Task accomplishment Table 1 reports the task accomplishment rate, which is found to vary across models and datasets. We find LLMs sometimes fail to adhere to the specified JSON format. This could happen when LLMs attempt to offer multiple possible answers or append additional text (e.g., repeated prompt footer or explanatory notes) to the response. Alternatively, the output could exceed the token limit, which is often the case when longer dialogue contexts are involved in the prompts.

In general, `Llama-3.1` and `Claude-3` outperform `Mistral-0.3` in following the instruction. All three LLMs achieve the highest task accomplishment rate on the WikiDialog dataset with

[4] Table 5 lists the 18 aspects covered by the FED metric.

[5] Table 7 shows the descriptions for all 24 aspects.

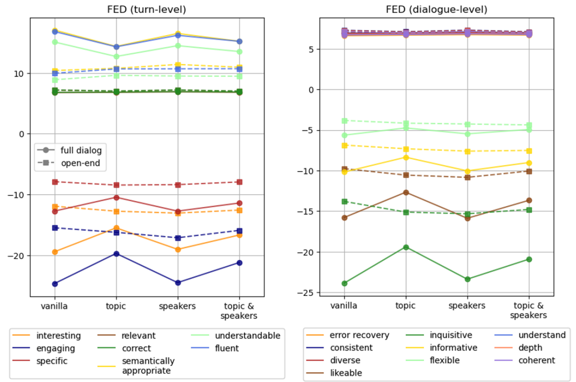


Figure 3: The different task performance of `Mistral-0.3` across three variables in prompt design: *topic*, *speakers*, and *open-end*.

synthetic dialogues. Although `Claude-3` generates mostly properly formatted responses on ReWIRED and WikiDialog, it struggles with ELI5, likely due to the longer dialogue contexts.[6]

Task performance Across the three datasets (Figures 14, 15, 16 in Appendix), the results for the dialogue completion task share similar trends with baseline experiments where we applied the test suites to the original data (App. A.3): LLMs tend to formulate explanatory turns well while interacting with the interlocutor rather poorly. Nevertheless, the divergence from baseline varies across datasets. On WikiDialog, the LLM-generated turns frequently reach similar scores to the original turn, while the gap is the widest for ReWIRED.

The three models exhibit inconsistent performance under different combinations of the variables. In many cases, the turns generated by `Claude-3` are rated similarly to the baseline, particularly on WikiDialog, but can sometimes be outperformed by the two open LLMs. Surprisingly, although `Llama-3` achieves better scores in most FED aspects in the trial run (App. B), `Mistral-0.3` often outperforms `Llama-3.1` in the full-size experiment. Even so, according to IXQUISITE, the turns filled-in by `Llama-3.1` share more similar linguistic features to the original turns, especially concerning text complexity. Regarding prompt design, *open-end* affects the output most drastically and mostly results in scores that deviate more from the baseline. In contrast, *topic & speakers* perturbs the output less, and the relative value compared to the baseline fluctuates (Fig. 17).

	ReWIRED	WikiDialog	ELI5-dialogues
by dataset	**specific** **flexible** **error recovery** coherence inquisitive	informative **consistent** **diverse** **error recovery** interesting	**consistent** depth **diverse** **flexible** **specific**
	`Mistral-0.3`	`Llama-3.1`	`Claude-3`
by model	**specific** **flexible** **inquisitive** likeable **depth**	**specific** **flexible** error recovery informative **inquisitive**	consistent **flexible** min. explanations diverse **depth**

Table 2: The five features with the lowest normalized Δf scores by dataset and by model. Features that appear more than once are highlighted in boldface.

4.2 Tuning filled-in turns

Prompting strategies Table 2 lists the worst-performing features across datasets and models, showing that the filled-in dialogue turns commonly perform worse in aspects, such as specificity, flexibility, and consistency.[7] Using ReWIRED as an example, the prompt for tuning outperforms the adapted original prompt in every FED aspect (Fig. 17). Therefore, the other datasets and models are later tuned only with the designated prompt, which also aligns better to the self-refine prompting framework (Madaan et al., 2023).

The tuned performance also varies across model and prompting variables. After instruction-tuning, the dialogue turns generated by `Llama-3.1` are often rated better than the baseline, whereas the gap between original and filled-in turn remains for `Mistral-0.3`. On the other hand, with the partially eliminated context in *open-end*, instruction-tuning introduces more drastic performance changes but also uncertainty in that the refined output can be rated worse more often.

Self-refine tuning In most scenarios[8], the tuned outputs achieve higher scores in most FED aspects, but in cases where LLMs tend to outperform human explanations (e.g., "semantically appropriate," "understandable," and "fluent"), the scores decrease. Such trend applies to `Mistral-0.3` and `Llama-3.1` in most settings, tuning `Claude-3` often worsens the results, e.g., WikiDialog with *topic & speakers & open-end* prompt (Fig. 20), and on ELI5 across both prompt variants (Fig. 21). Overall, `Claude-3` often achieves the best scores in

[6]For each dataset (ReWIRED, WikiDialog, and ELI5-dialogues), a sample instance and corresponding model output is provided in App. A.2.

[7]Although IXQUISITE scores do not directly imply dialogue quality, we do not exclude them in feature selection because the FED aspects remain more dominant.

[8]Charts are provided in App. D.

the dialogue completion task, limiting its potential of further refinements. In comparison, while `Llama-3.1` usually performs worse on the task, instruction-tuning substantially improves its output, even under scenarios where the initial output already appears promising (Fig. 21).

A similar trend can be observed on a dataset scale: Substantial improvements following tuning are more likely when the gap is larger between task performance and the baseline. With the same vanilla prompt, the output of ELI5 performs rather consistently before and after tuning in comparison to the other two datasets.

5 Discussion

5.1 LLM output of dialogue completion

For the dialogue completion task, the results show that the LLMs' outputs align more closely with the provided dialogue context than with additional information appended to the task description. Leaving out subsequent turns induces more space for potential output than other explored prompt modifications. At every stage, applying *open-end* usually leads to poor results. This trend confirms the gap between how human speakers and LLMs address explanations, as models usually fail to perform robustly under the reduced dialogue context: Examples of *open-end* (App. A.2) further show that LLMs often incorporate different details in an explanation, which leads to misaligned content between the model-generated turn and the subsequent ones. However, the written turns in ELI5 are an exception, suggesting that LLMs can come up with explanatory turns that approximate the human ones within their output space.

Turn length of the output is also sensitive to the dialogue context. The trend is particularly relevant to the task accomplishment rate of `Claude-3` on ELI5, which is the lowest overall (Table 1). Many model-generated turns are discontinued because the token limit is reached, which eventually results in uncompleted JSON objects. Since the written dialogues tend to feature longer turns, models may also respond with longer ones.

In comparison, **task description** makes little difference to the performance. Specifying dialogue topic or expertise level of speakers may sometimes introduce slight performance enhancement, such as the FED aspects shown in Figure 14, but most of the time, prompt variables *topic* and *speakers* hardly affect FED and IXQUISITE scores.

5.2 Dialogue quality

Thanks to their sources, all datasets exhibit distinct characteristics according to FED and IXQUISITE (App. A.3), revealing how the metrics rate human-spoken (ReWIRED), synthetic (WikiDialog), and human-written dialogues (ELI5).

Although ELI5 outperforms the other two datasets in most FED aspects, when looking at turn-level and dialogue-level assessments, WikiDialog achieves better scores on **turn level** ("fluent", "understandable", "semantically appropriate") where ELI5 turns are reported to perform even worse than the ReWIRED ones. Such tendency reflects the characteristics of written text: Constructed from Wikipedia entries, explanatory turns in WikiDialog are expected to be more sophisticated in language use. Nevertheless, the written turns from ELI5 obtain lower scores than the ReWIRED transcripts, because field experts (ReWIRED) presumably explain a concept better than random Internet users.

In comparison to the turn-level criteria, ELI5 reaches the best scores on all **dialogue-level** FED aspects. This proves that human conversations are more natural as a whole and the dialogue flow is considered superior to that of synthetic dialogues. The score differences to ReWIRED and WikiDialog further reflect the nuances between colloquial and written language, especially in aspects where the scores of ReWIRED deviate more heavily from those of WikiDialog and ELI5 ("understanding", "coherent", "topic depth"). Written texts tend to obtain higher scores on delivering information, while interactivity of transcripts would usually be rated better. Nevertheless, the IXQUISITE metric says that written texts involve significantly more complex language use ("lexical complexity", "readability level") and less repetition ("minimal explanations"), while this difference is smaller between written and spoken dialogues. These findings confirm the complexity of explanatory dialogues and the challenges for evaluation, as linguistic features do not directly determine dialogue quality but instead contribute to how speakers could possibly engage with and perceive a conversation.

Despite similar characteristics according to IXQUISITE, human-written dialogues outperform the spoken ones in all FED aspects. Speaking is typically regarded as the most fundamental form of conversation, so this counterintuitive misalignment raises concerns about biases in FED towards written language and its use for dialogue evaluation.

Evaluated Aspect	Mistral-0.3	Llama-3.1	Claude-3
Relevant	-33.3	-37.5	-51.5
Interesting	-40.6	-51.0	-33.1
Fluent	46.1	53.4	40.3
Correct	-32.6	-36.8	-51.1
Understandable	45.0	52.4	37.8
Semantically appropriate	38.2	49.0	32.8
Specific	-45.7	-50.6	-35.2
Engaging	-36.5	-47.4	-26.6
Likeable	-37.6	-46.6	-25.4
Understand	-33.3	-37.6	-51.6
Coherent	-33.3	-37.6	-51.6
Depth	-32.6	-36.5	-50.9
Flexible	-13.7	-19.7	-12.8
Diverse	-31.6	-35.2	-50.2
Inquisitive	-42.4	-50.1	-36.7
Error recovery	-34.6	-39.4	-52.6
Informative	-33.2	-44.5	-38.1
Consistent	-31.0	-34.5	-49.8
Minimal explanations	-16.9	-15.7	-20.3
Lexical complexity	48.2	7.3	38.5
Synonym density	-3.7	0.2	1.4
Adaptation	-3.6	-10.5	8.9
Reading grade	14.5	1.4	1.0
Coherence	-1.8	1.3	3.1

Table 3: Normalized evaluation results in percentage (%) of each model on all datasets using the vanilla prompt. The best scores for each aspect are highlighted.

5.3 Model-generated and human turns

LLM-generated turns can function as explanations and yet often fail to put interactivity into practice, confirming the gap between synthetic and human dialogues (Dai et al., 2022; Stacey et al., 2024). Models perform the task better on dialogues more closely aligned with the impression that human readers intuitively associate LLM output, such as being grammatically correct and elaborate. Overall, the task output for ReWIRED deviates from the baseline the most according to FED, while the gap is also confirmed by IXQUISITE in that model-generated turns tend to feature more complex utterances. Yet this does not mean LLMs work well with all the synthetic data; WikiDialog does not necessarily feature the characteristics of model-generated dialogues. Likewise, the detailed expressions in ELI5 may contribute to its outstanding score in various aspects.

LLMs are efficient in organizing information into sentences that make sense, even within a dialogue format, but perform poorly in providing the interlocutor with sufficient information (“topic depth”, “informativeness”). Insufficient specificity of generated dialogues has also been pointed out by Wang et al. (2024). Across the evaluated aspects, the three models exhibit divergent performance (Table 3): While `Claude-3` tends to generate succinct output, `Mistral-0.3` often provides detailed explanations (App. A.2). Although seemingly better at first glance (Figures 14 & 16), `Claude-3` doesn’t outperform the other open LLMs, but it is rated higher at interaction (“engaging”, “likeable”), which is commonly considered challenging for dialogue systems, and especially for `Llama-3.1`. Its poor interactivity is related to the verbosity of explanations, which may fail to address the query implied by the dialogue context (Li et al., 2024).

The IXQUISITE metric tells us that `Mistral` favors more complex expressions, while `Claude-3` tends to employ slightly more conjunctions. Alongside FED, we observe a slight correlation between lexical complexity and informativeness, and a better interactivity through the use of plain language and connective words. The scores supply the FED suite with a bridge between the pragmatic outcomes and quantifiable linguistic features.

5.4 Self-refined filled-in turns

LLMs are capable of improving their own output for the dialogue completion task. Our instruction-tuning strategies, inspired by Madaan et al. (2023), prove that these improvements are independent from dialogue history and can be achieved simply with zero-shot instructions.

Although the original output and the refined output often incorporate analogous contents (App. A.2), the different evaluation outcomes confirm that explanations rely not only on content but also on expression (Wachsmuth and Alshomary, 2022), and that how a concept evolves is equally important. Figures 19 to 21 show that `Llama-3.1` consistently yields substantial improvements for the FED aspects, which is not always the case for the other two models. While the improvement of `Mistral` beats `Claude`, both models perform better on ReWIRED than the other two datasets, where instruction-tuning sometimes decreased the scores. This tendency could be related to the task performance, in which model-generated turns of ReWIRED often show the largest gap from the baselines, leaving more room for improvement. The prompt variable *open-end* limits the models’ improvement, as the short context introduces more uncertainty, even with additional instructions.

Aspects measured by FED are correlated, since a reference utterance can be shared by multiple aspects and contribute to their respective scores simultaneously. As a result, instructions focusing on certain aspects in fact introduce an overall improvement. The overlap between evaluated aspects could also be observed from the sensitivity to prompt variables (Fig. 4), as different FED aspects could exhibit similar tendencies under such perturbation.

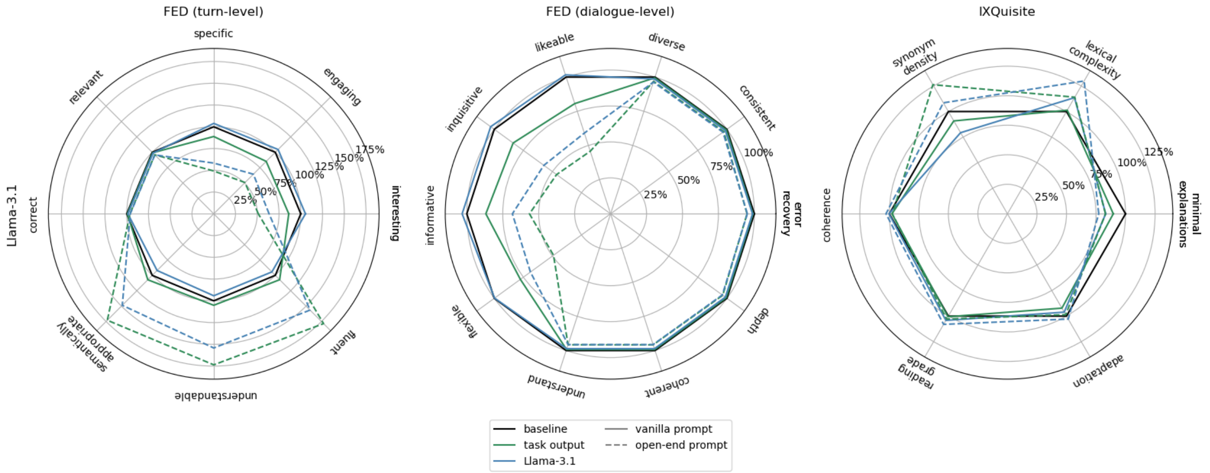


Figure 4: The evaluation results of instruct-tuned turns of `Llama-3.1` on the ReWIRED dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the blue lines indicate the tuned output of `Llama-3.1`. Solid and dashed lines respectively represent vanilla and *open-end* prompt.

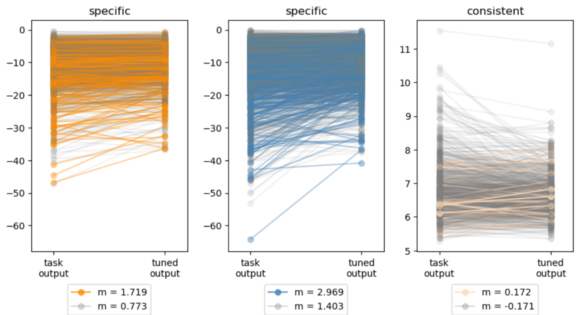


Figure 5: Instance-level score change after instruction-tuning (l: `Mistral-0.3`, c: `Llama-3.1`, r: `Claude-3`), w.r.t. their worst-performing feature (Table 2) on ReWIRED. Changes in scores are represented by slope m. Model-specific colors highlight instances where the given aspect is described in the instruction.

On instance level, specifying aspects in the prompt results in substantial score increases compared to the overall improvement. For Figure 5, we observe a steeper gradient where the given aspect is described. From all combinations of prompt variables on ReWIRED, we find the prompted instruction to correlate with the quality on the corresponding evaluated aspect of explanatory dialogues.

6 Conclusion

In this work, we conducted a comprehensive study of the dialogue completion task in three distinct explanation domains: Human spoken, human written, and synthetic. We are the first to explore dialogue completion – as proposed in Dai et al. (2022) and picked up by Li et al. (2023); Wang et al. (2024); Li et al. (2024), i.a. – for explanatory dialogue across multiple domains. We found LLMs capable of efficiently formulating knowledge or information for explanatory dialogues, but they often remain inadequate to replace human explanations owing to the poor interactivity. The output space of the dialogue completion task highly depends on how the dialogue context is established. Although the word choices in the prompt have a small effect, the changes are often negligible in comparison to those introduced by trimming dialogue context. In contrast, a reduced context significantly lowers explanation quality on both turn and dialogue level.

According to two reference-free test suites, LLMs tend to suffer from the interplay with the interlocutor, especially while taking the entire conversational context into account rather than a single turn. LLMs fail to achieve the baseline's performance on human-spoken conversations, likely owing to the wider disparity between colloquial and written language, as well as the differences between model-generated and human-produced contents. The corpus with human-written explanations from online forums was rated highest, while ReWIRED is constantly considered as the lowest quality, presumably due to editing and intrinsic biases towards written dialogues. Despite the gap between the original and filled-in turns, LLMs could refine their own task output through prompt-based methods and metric-based feedback, where smaller ones like `Llama-3.1` especially profit from. Our analysis of the worst-performing features reveals that specific aspects of underperformance still vary across models and tasks.

Limitations

The **datasets** adopted in this work are rather small, especially after resampling WikiDialog and ELI5-dialogues to align with the size of ReWIRED. Considering the task accomplishment rate, the final amount of data points per batch sometimes falls below 200, which could limit the robustness of the findings and increase susceptibility to sampling error. In addition to the limited data points, biases could originate from the minor flaws in the experimental pipeline. For example, although the IXQUISITE aspect "synonym density" is left out for WikiDialog and ELI5-dialogues, the aspect continues to exist in instruction-tuning and occasionally becomes one of the worst features.

The FED metrics' **preference for written dialogues** could possibly relate to the pretrained `DialoGPT` model. Although conversations can be transcribed, written chats are presumably more readily available as data. If the model was trained more on dialogues in form of messages or texts, it would more likely predict written language as its output, which eventually leads to higher scores. Another potential cause is that the *5 Levels* videos (source of the (Re)WIRED dataset) were heavily edited. The production process may create gaps between dialogue turns, causing lower scores. Nevertheless, such preference doesn't necessarily imply poor reliability of the scores; considering that applications such as dialogue-oriented NLE commonly concern written dialogues, the metric is still considered robust while highlighting the gap between data sources.

References

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

Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. Modeling the quality of dialogical explanations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11523–11536. ELRA and ICCL.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

David Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15607–15631. Association for Computational Linguistics.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

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@ACL 2023, Toronto, Canada, 13 July 2023*, pages 528–538. Association for Computational Linguistics.

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

Nils Feldhus, Aliki Anagnostopoulou, Qianli Wang, Milad Alshomary, Henning Wachsmuth, Daniel Sonntag, and Sebastian Möller. 2024. Towards modeling and evaluating instructional explanations in teacher-student dialogues. In *Proceedings of the 2024 International Conference on Information Technology for Social Good, GoodIT 2024, Bremen, Germany, September 4-6, 2024*, pages 225–230. ACM.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. Interrolang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5399–5421. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. 2024. Leveraging llms for dialogue quality measurement. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL*

2024, Mexico City, Mexico, June 16-21, 2024, pages 359–367. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. Tigerscore: Towards building explainable metric for all text generation tasks. *Trans. Mach. Learn. Res.*, 2024.

Soonwoo Kwon, Sojung Kim, Minju Park, Seunghyun Lee, and Kyuseok Kim. 2024. BIPED: Pedagogically informed tutoring system for ESL education. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3389–3414, Bangkok, Thailand. Association for Computational Linguistics.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *HCAI @ NeurIPS 2022*.

Grace Li, Milad Alshomary, and Smaranda Muresan. 2024. "is chatgpt a better explainer than my professor?": Evaluating the explanation capabilities of llms in conversation compared to a human baseline. *CoRR*, abs/2406.18512.

Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. AutoConv: Automatically generating information-seeking conversations with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

Longxuan Ma, Ziyu Zhuang, Weinan Zhang, Mingda Li, and Ting Liu. 2022. Self-eval: Self-supervised fine-grained dialogue evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 485–495. International Committee on Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Shikib Mehri and Maxine Eskénazi. 2020b. USR: an unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 681–707. Association for Computational Linguistics.

John Mendonça, Isabel Trancoso, and Alon Lavie. 2024. Soda-eval: Open-domain dialogue evaluation in the age of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11687–11708, Miami, Florida, USA. Association for Computational Linguistics.

Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: a systematic survey. *Artif. Intell. Rev.*, 56(4):3055–3155.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4932–4942. Association for Computational Linguistics.

Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor, and Anders Johannsen. 2024. LUCID: LLM-generated utterances for complex and interesting dialogues. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 56–74, Mexico City, Mexico. Association for Computational Linguistics.

Anaïs Tack and Chris Piech. 2022. The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022, Durham, UK, July 24-27, 2022*. International Educational Data Mining Society.

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 344–354. International Committee on Computational Linguistics.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5967–5994. Association for Computational Linguistics.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. Dynaeval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5676–5689. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.

A Data

A.1 Datasets

The datasets adopted in the experiments are described as below:

WIRED / ReWIRED. Focusing on explanatory dialogues on scientific topics, the WIRED dataset (Wachsmuth and Alshomary, 2022) transcribes human conversations from 65 edited video clips in which 13 high-level subjects are explained by a field expert to different target audience. For each topic, the dataset includes five distinct expertise levels of explainee: child, teenager, undergraduate student, graduate student, and colleague, i.e., another field expert. Similar concepts could always develop differently in the way of being phrased and argued (Figure 6). Moreover, the dataset features turn-level labels that highlight interactions categorized into dialogue acts, and explanation acts. The expanded ReWIRED corpus in Feldhus et al. (2024) adds 65 more transcripts and features token-level annotations that incorporate teaching acts.

WikiDialog. Dai et al. (2022) built this dataset from Wikipedia texts through "inpainting" the missing part of a dialogue; to be specific, raw texts from the documents were segmented into explainer turns that interact with an imagined explainee agent, and the explainee turns were then filled-in with the predicted output from a designated "inpainter" model. The validation set originally consists of 113,820 synthetic dialogues. We remove the dialogues that are shorter than 10 turns; moreover, since Wikipedia pages frequently focus on specific individuals, we apply a rough NER filter to dialogue topics with NLTK toolkits for removing those dialogues, resulting in a total of 42,573 dialogues.

ELI5-dialogues. This dataset by Alshomary et al. (2024) extracts around 400 conversations from the *Explain Like I'm Five* subreddit, where users ask questions and request understandable explanations. ELI5 demonstrates how explanations could be delivered in written form, which better resembles the interaction with LLMs. The corpus contains 2,650 annotated data instances. Considering written texts can sometimes be longer than transcribed oral conversations, we remove the dialogues containing lengthy turns that are over 200 tokens. This slightly reduces the corpus size to 2,056 data instances.

A.2 Sample data points

Table 9 to 11 include one data instance per dataset, showing the retrieved explanatory dialogue context, the original turn, and the output of each model turns under the vanilla prompt.

Table 12 demonstrates a comparison between task output with and without *open-end* prompt variable, which is found to impact model output the most. Furthermore, Table 13 sheds light to how instruction-tuning may alter the output while phrasing mostly the same contents. In both tables, distinction between the two output variants are manually highlighted. In light of the metric scores in the evaluation results, the misalignment of these highlighted details may reflect the qualitative distinction that ultimately contribute to nuances in the quantitative analyses.

A.3 Dataset quality (baseline)

Table 4 presents the evaluation results of the original turns per dataset as a baseline. In almost all the FED aspects, ELI5 achieves the best score, while WikiDialog occasionally outperforms the other two datasets. On the other hand, the IXQUISITE metric reveals that the WIRED dataset scores slightly higher in "adaptation" and "coherence", WikiDialog receives high marks in some aspects such as "lexical complexity". While the FED metric provides a qualitative assessment for each aspect, the IXQUISITE scores do not necessarily imply dialogue quality; instead, they offer mostly descriptions of the form and function of the language features. For example, complicated diction and syntax may appear more suitable or common to certain interlocutors, but the level of lexical complexity itself remains neutral.

B Trial run

We conduct a trial run on the original WIRED dataset with `Mistral-0.3 7B` (Jiang et al., 2023) and `Llama-3 8B` (Grattafiori et al., 2024) to decide the variables to be included in the experiments. With l set to 60 tokens and w set to 2 turns, the trial batch outputs in total 2,632 segmented dialogues filled-in by LLMs. Among all data points, the model-generated turns contain 1.59 sentences with a length of 30.39 tokens on average.

Evaluated with FED and IXQUISITE (Figure 8), the trial run hints at the performance difference carried out by the four prompt variables (Figure 2). The variable *footer w/ context* is left out because LLMs follow the instructed output format poorly, presumably owing to the longer and repetitive input. Under the remaining eight scenarios derived from combination of *topic*, *speakers*, and *open-end*, `Mistral-0.3` (Figure 10) and `Llama-3` (Figure 11) are rated similarly by the two test suites: Applying *open-end* drastically affects model performance, while *topic* and *speakers* introduce less changes to the scores.

The outcome of the trial run introduces adjustments to the subsequent experiment on the main task: First, considering the output length, minimum length (l) for instance building is reduced from 60 to 30. Second, regarding prompt structure, *footer w/ context* is abolished, while *topic* and *speakers* are combined into one variable in the following experiments.

Evaluated Aspect	ReWIRED	WikiDialog	ELI5
Relevant	6.92	7.56	**7.73**
Interesting	-11.13	-15.18	**-9.15**
Fluent	9.90	**12.73**	8.17
Correct	6.94	7.61	**7.76**
Understandable	9.06	**11.90**	7.63
Semantically appr.	10.44	**13.11**	8.45
Specific	-7.15	-11.00	**-6.02**
Engaging	-14.14	-18.46	**-11.63**
Likeable	-8.33	-12.57	**-6.74**
Understanding	6.91	7.56	**7.74**
Coherent	6.90	7.55	**7.72**
Topic depth	6.95	7.62	**7.76**
Flexible	-3.37	-4.66	**-3.11**
Diverse	6.99	7.67	**7.79**
Inquisitive	-12.44	-18.67	**-9.63**
Error recovery	6.84	7.46	**7.68**
Informative	-6.73	-5.63	**-4.88**
Consistent	7.02	7.71	**7.82**
Minimal explanations	0.16	**0.75**	0.18
Lexical complexity	1.15	**2.33**	1.44
Synonym density	**0.09**	0.00*	0.00*
Adaptation	**0.24**	0.21	**0.24**
Readability level	0.46	**0.63**	0.45
Coherence	**0.04**	0.03	0.03

Table 4: The differences between dialogues from the three datasets according to the two selected test suites. The best scores for each aspect is highlighted in boldface. The starred zeros result from limitations of the datasets, as described in App. C.

C Edge cases

Among the IXQUISITE aspects, "synonym density" does not apply to ELI5-dialogues because the dataset does not contain topic-related keywords for each dialogue, whereas the test suite requires at least one to derive synonyms. For the WikiDialog dataset, where page titles are applied, the feature hardly captures anything because Wikipedia pages are usually proper nouns or named entities, which can hardly be substituted by other words.

D Supplementary results

Table 8 extends the results in Table 1 and describes the amounts of data instances where the LLM successfully accomplish the dialogue completion task. Task accomplishment rate is found to vary across tasks and prompt variables.

As an extension of Table 2, Figure 12 and 13 demonstrate the counts for worst-performing features that are retrieved for instruction-tuning.

Across two prompt variants, the results are respectively visualized for each dataset: ReWIRED

Turn-level	Dialogue-level
Relevant	Likeable
Interesting	Understanding
Fluent	Coherent
Correct	Topic depth
Understandable	Flexible
Semantically appropriate	Diverse
Specific	Inquisitive
Engaging	Error recovery
	Informative
	Consistent

Table 5: The 8 turn-level and 10 dialogue-level aspects measured by FED metric (Mehri and Eskénazi, 2020a), ordered by importance reported in the original paper.

(Figure 19), WikiDialog (Figure 20), and ELI5-dialogues (Figure 21). In these radar charts, baselines are represented by regular polygons defined as 100% in each dimension. The colored polygons denote the proportional scores relative to the baseline for the task (green) and tuned (model-specific color) output. A score higher than the baseline moves the dimension closer to the edge, while a lower score brings it closer to the center.

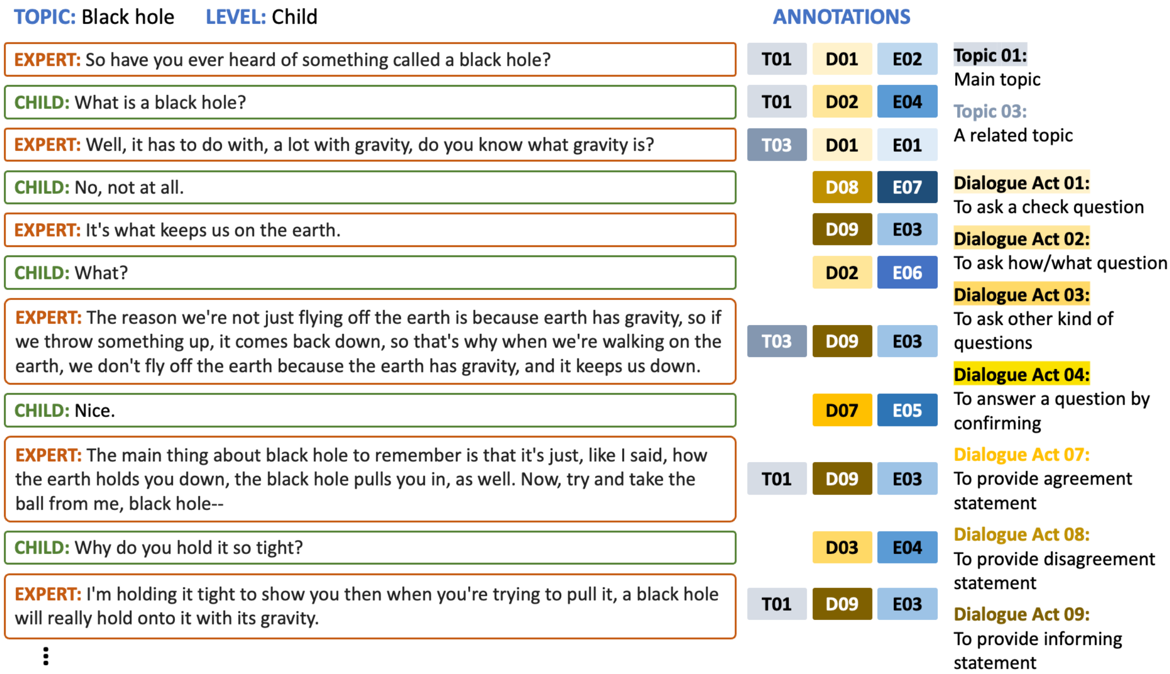


Figure 6: An excerpt of a teacher-child explanatory dialogue from the WIRED dataset (Wachsmuth and Alshomary, 2022), exemplifying explaining the same topic to different explainees (see Figure 7). The figure also highlights the information provided in the corpus: topic, explainee's expertise level, and turn-level annotations, along with their contents on the right.

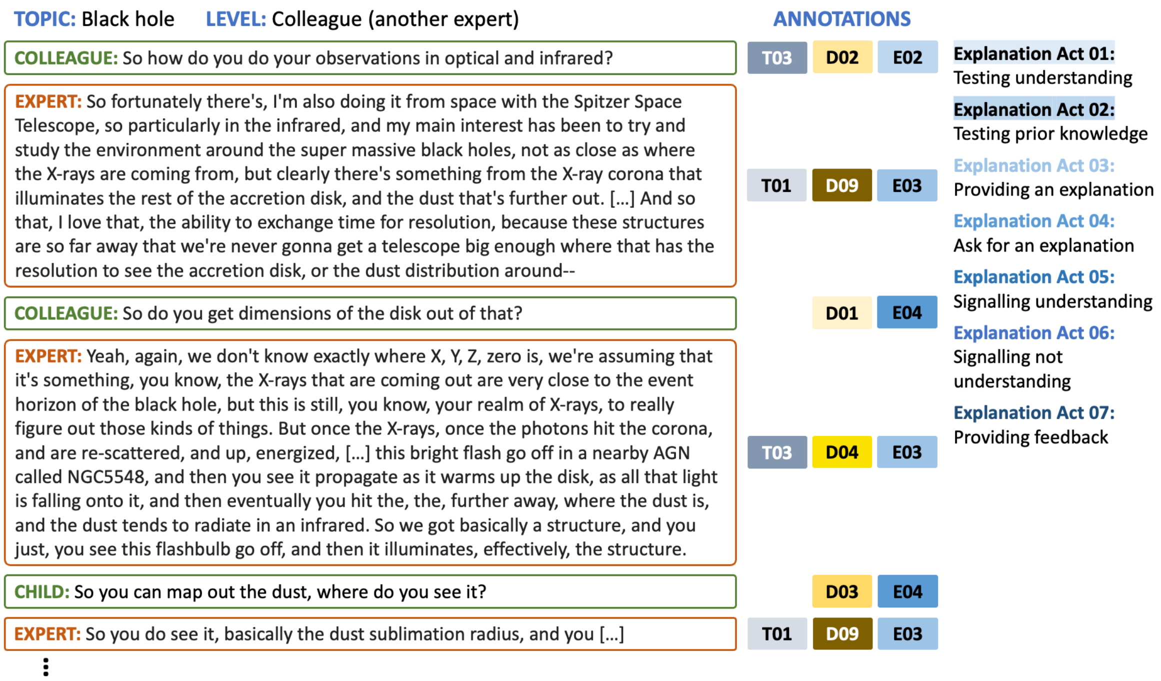


Figure 7: An excerpt of a teacher-colleague explanatory dialogue from the WIRED dataset (Wachsmuth and Alshomary, 2022).

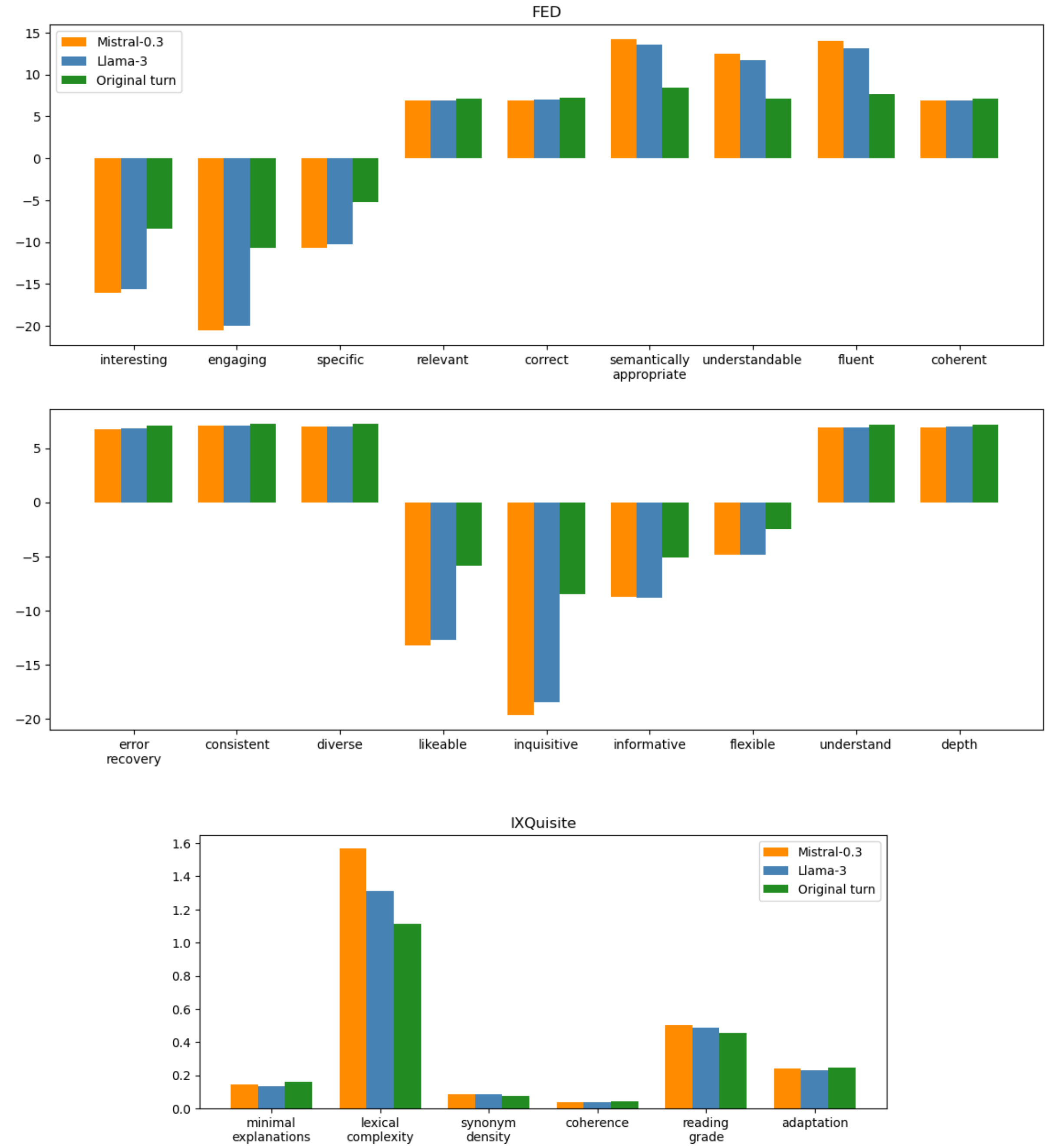


Figure 8: Raw metric scores of the LLM-generated and the original human turns measured by FED and IXQUISITE.

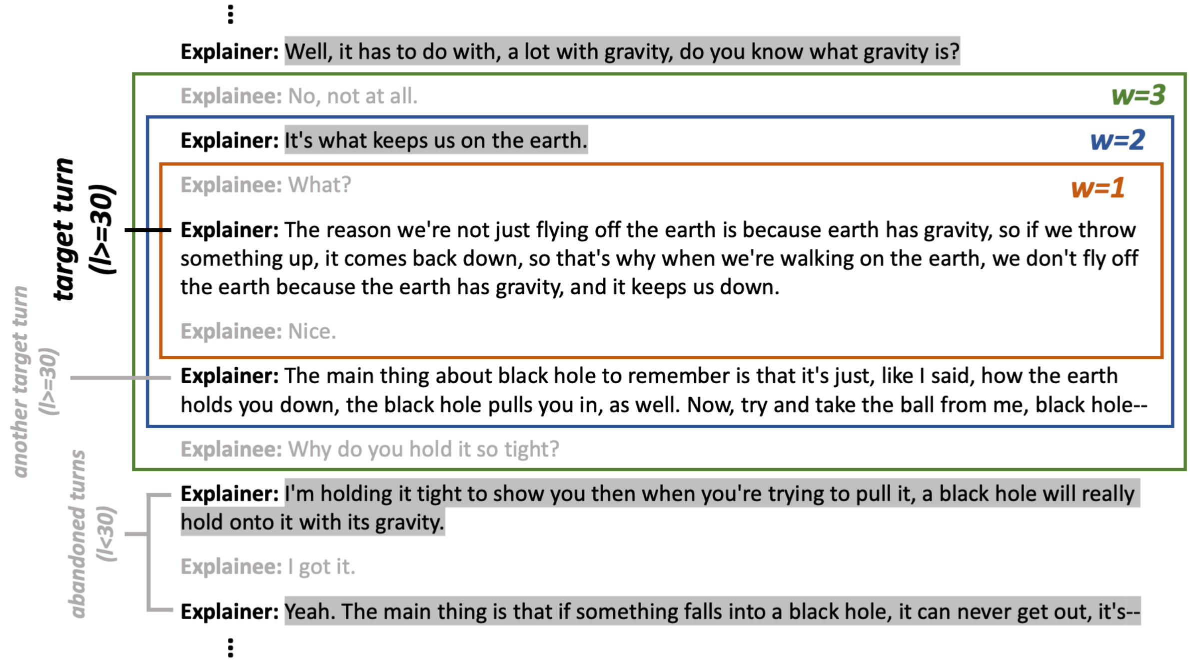


Figure 9: Example from the WIRED dataset of how data instances are built from a turn and its surrounding dialogue context. Minimum turn length l is set to 30 here.

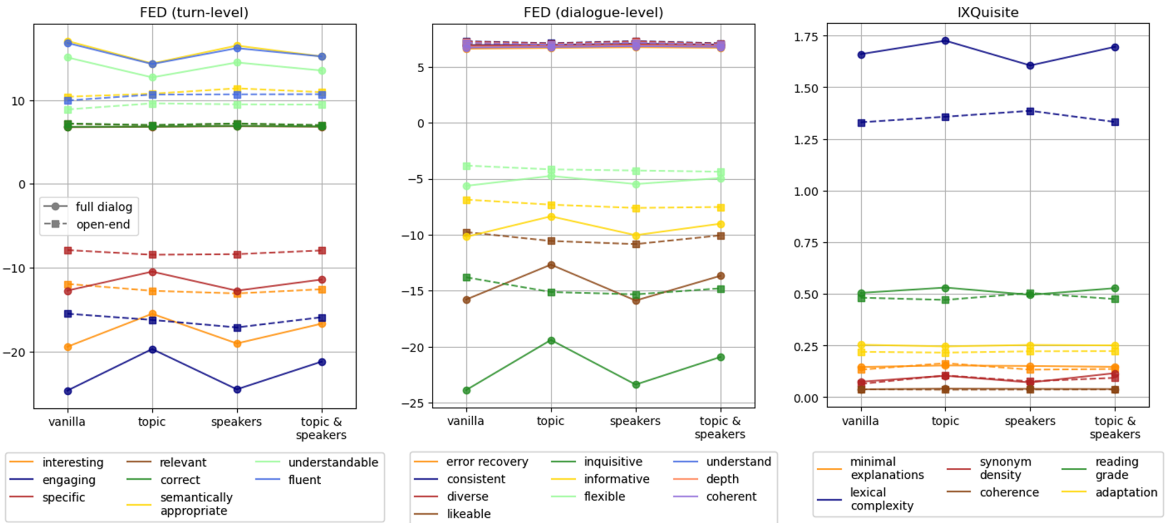


Figure 10: The different task performance of `Mistral-0.3` across three variables in prompt design: *topic*, *speakers*, and *open-end*.

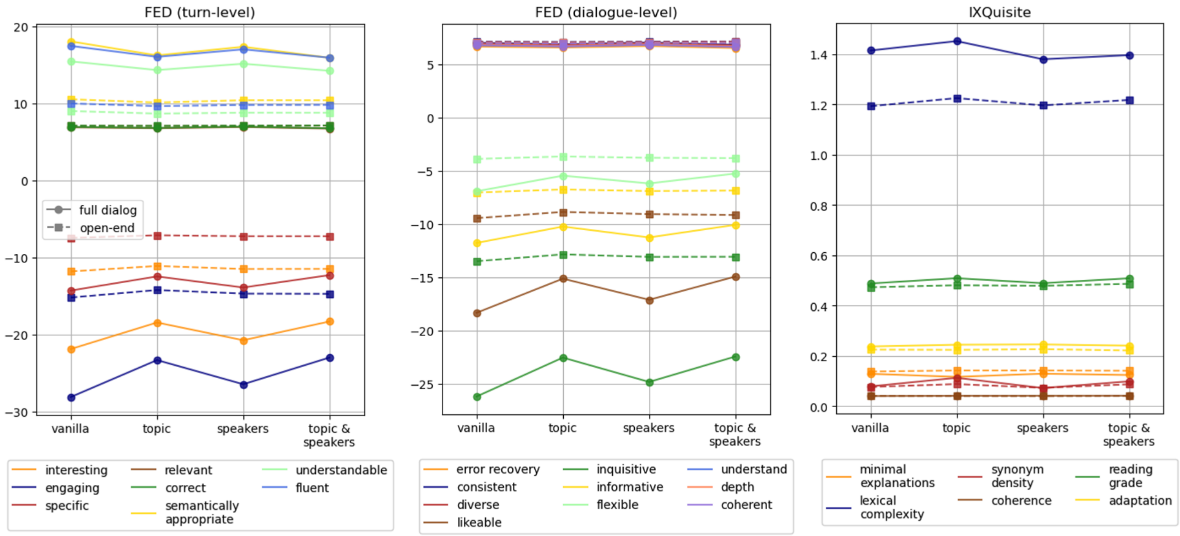


Figure 11: The different task performance of `Llama-3` across three variables in prompt design: *topic*, *speakers*, and *open-end*.

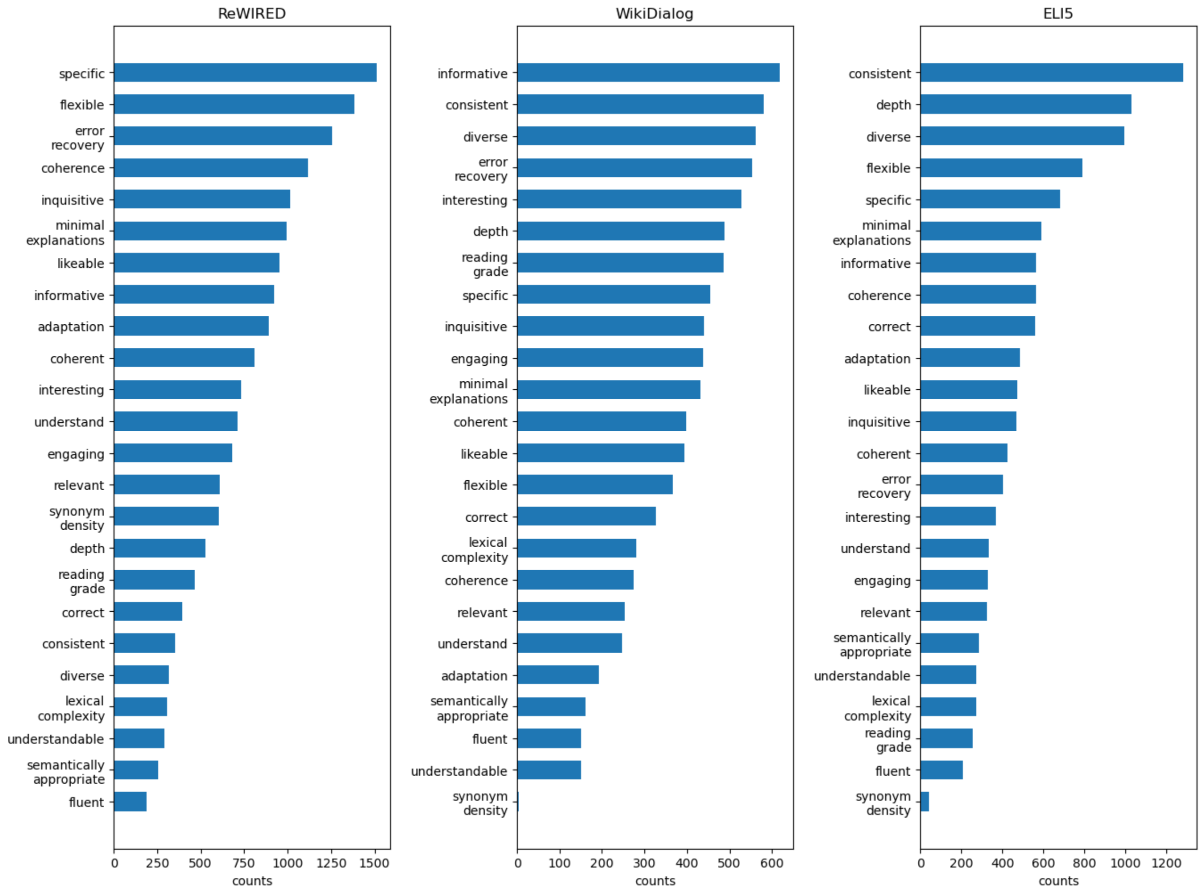


Figure 12: The worst performing features after normalization for each dataset.

Category	Measure
Minimal explanations	Frequency of named entities
Lexical complexity	Frequency of difficult words
Synonym density	Frequency of synonyms for the relevant terms
Adaptation	Inverse frequency of synonyms
Readability level	Flesch-Kincaid Grade level
Coherence	Frequency of conjunctions and connective words

Table 6: The 6 numerical measures in IXQUISITE (Feldhus et al., 2024) adopted in this work. Correlation to teaching model from the original test suite is left out because of the lack of annotations.

Aspect	Neg. Description	Pos. Description
Relevant	appear irrelevant	topic-relevant
Interesting	appear boring	interesting
Fluent	be poorly phrased	fluent
Correct	misunderstand the conversational context	correct
Understandable	be hardly understandable	understandable
Semantically appr.	make little sense	semantically appr.
Specific	appear out of scope	topic-specific
Engaging	appear unappealing	engaging
Likeable	appear unfriendly	likeable
Understanding	misunderstand the other speaker	perceptive
Coherent	deviate from the topic	coherent
Topic depth	appear superficial	in-depth
Flexible	adapt poorly to the conversation flow	flexible
Diverse	include too much repetition	lexically diverse
Inquisitive	appear indifferent	inquisitive
Error recovery	appear erroneous	self-corrective
Informative	provide too little information	informative
Consistent	disagree with previous utterances	consistent
Minimal explanations	mention too many named entities	accessible
Lexical complexity	incorporate difficult word usage	colloquial
Synonym density	paraphrase too little	lexically diverse
Adaptation	emphasize the same things too much	adaptive
Readability level	appear too hard to understand	plain
Coherence	introduce poor dialogue flow	coherent

Table 7: The evaluated aspects and their corresponding negative and positive description for instruction-tuning. The usage of both descriptions aims at narrowing down the semantic space of a single phrase: The FED metrics (Mehri and Eskénazi, 2020a) score dialogues based on the likelihood of responding with certain utterances, and these utterances can sometimes be shared by several aspects. For example, *That makes no sense!* contributes to the negative scores of "semantically appropriate," "understandable," and "fluent" simultaneously, implying that the measurement of these aspects is not discrete but mutually relevant. The phrasing of the descriptions takes into account of these source contents and further make syntactic adaptations to ensure the grammatical correctness of the later prompted instructions, resulting in all the negative descriptions being verb phrases and positive descriptions are adjective. Furthermore, although many positive descriptions simply inherit the name of the aspect, some of them are rephrased to appear appropriate in the sentence structure of the prompt.

Model	**Dataset**	vanilla	*topic, speakers*	*open-end*	*topic, speakers, open-end*
Mistral-0.3	ReWIRED	293	346	471	480
	WikiDialog	133	200	260	263
	ELI5-dialogues	234	212	411	415
Llama-3.1	ReWIRED	521	540	500	522
	WikiDialog	275	274	266	274
	ELI5-dialogues	473	438	446	433
Claude-3	ReWIRED	611	595	451	440
	WikiDialog	297	267	239	181
	ELI5-dialogues	310	290	168	169

Table 8: Task accomplishment counts across models, datasets, and prompt variables.

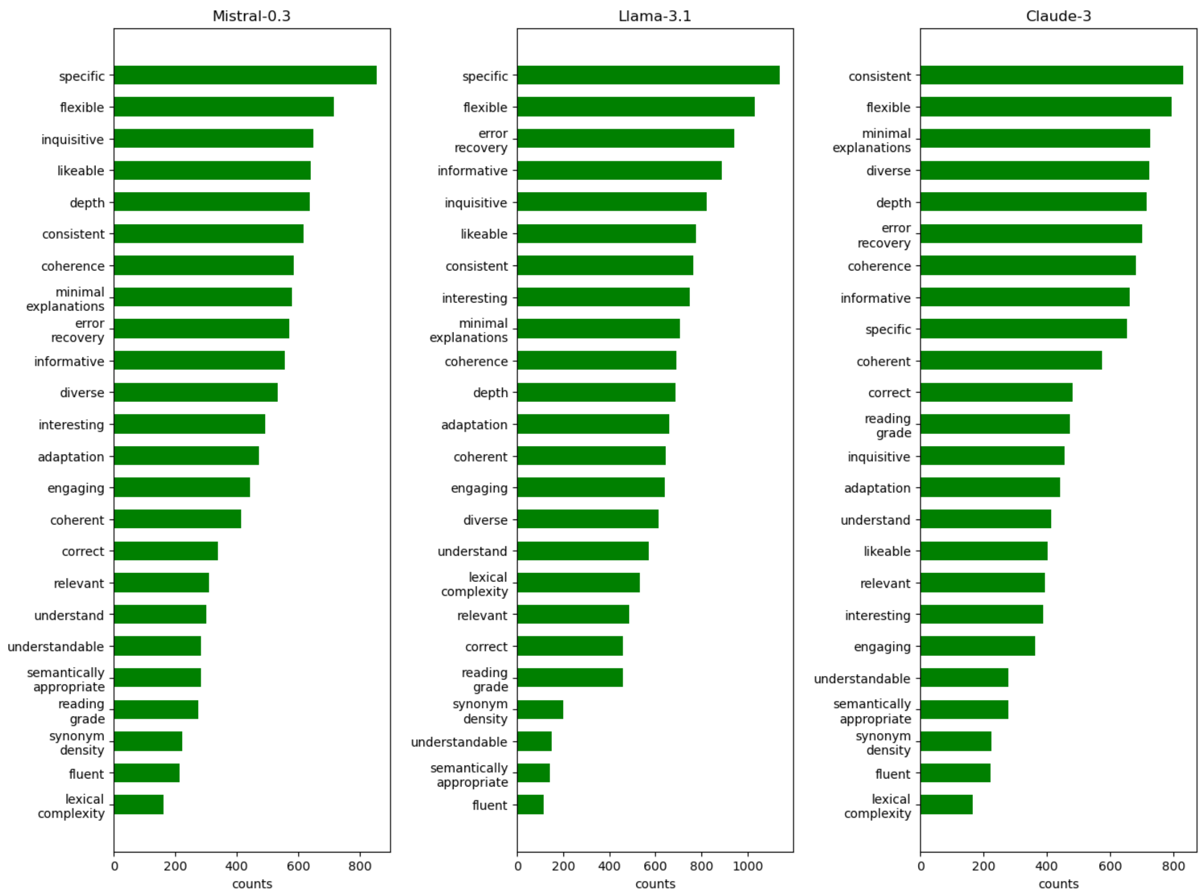


Figure 13: The worst performing features after normalization for each LLM.

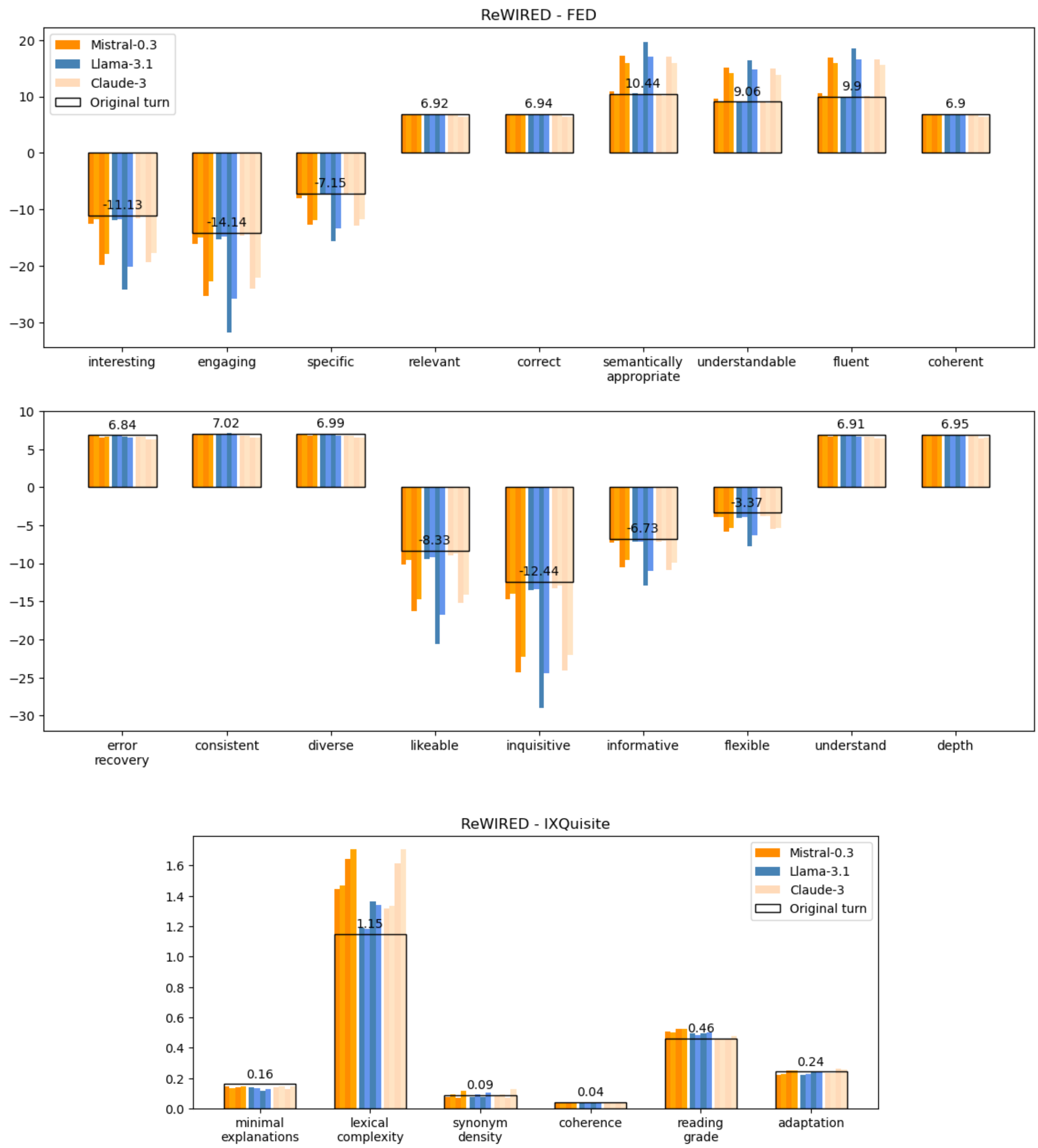


Figure 14: The evaluation results of filled-in turns of the ReWIRED dataset. From left to right, the four bars for each model represent the prompt variables: vanilla, *topic & speakers*, *open-end*, and *topic & speakers & open-end*.

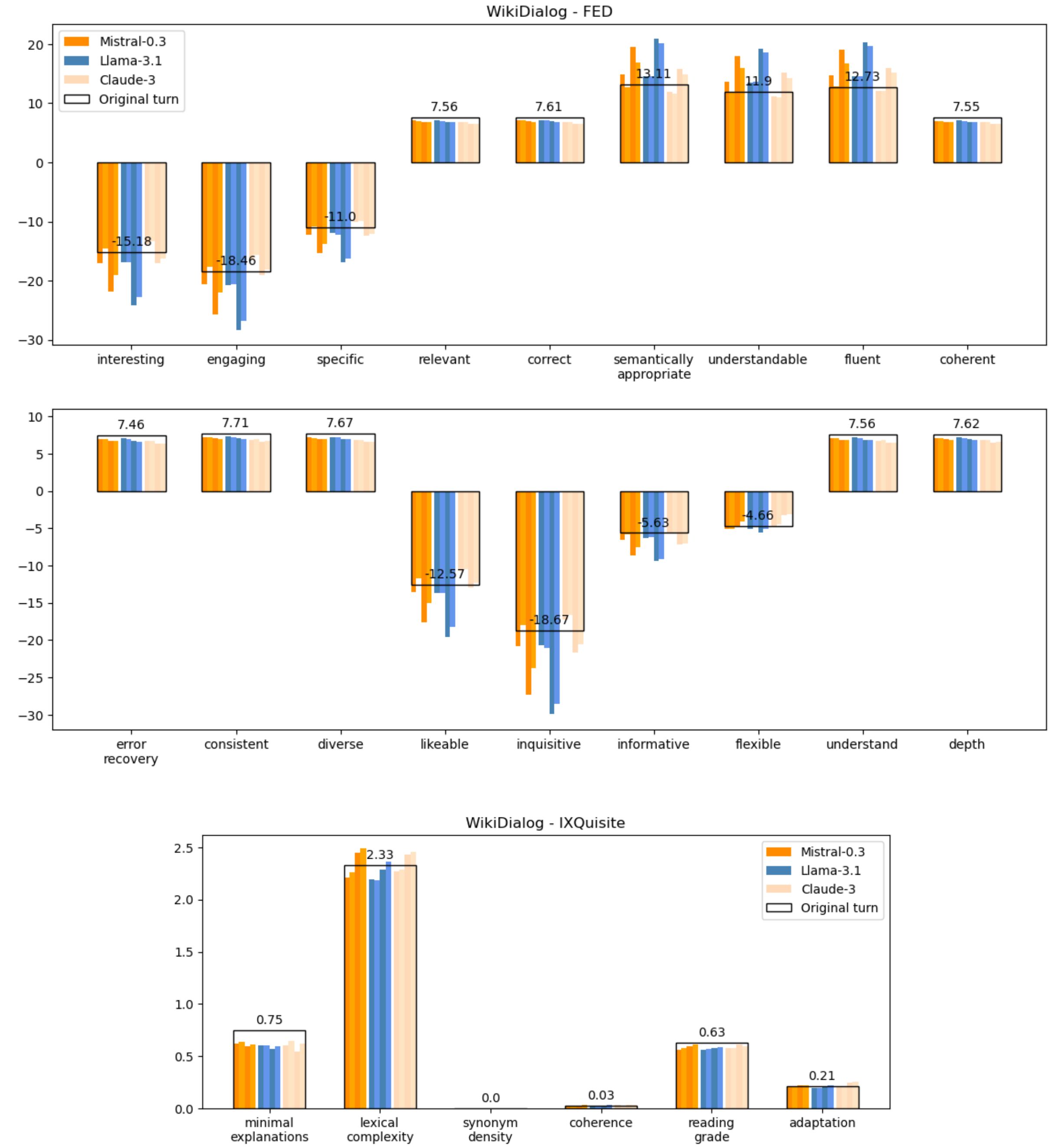


Figure 15: The evaluation results of filled-in turns of the WikiDialog dataset. From left to right, the four bars for each model represent the prompt variables: vanilla, *topic & speakers*, *open-end*, and *topic & speakers & open-end*. "Synonym density" is zero because page titles could hardly be paraphrased, as explained in Appendix A.3.

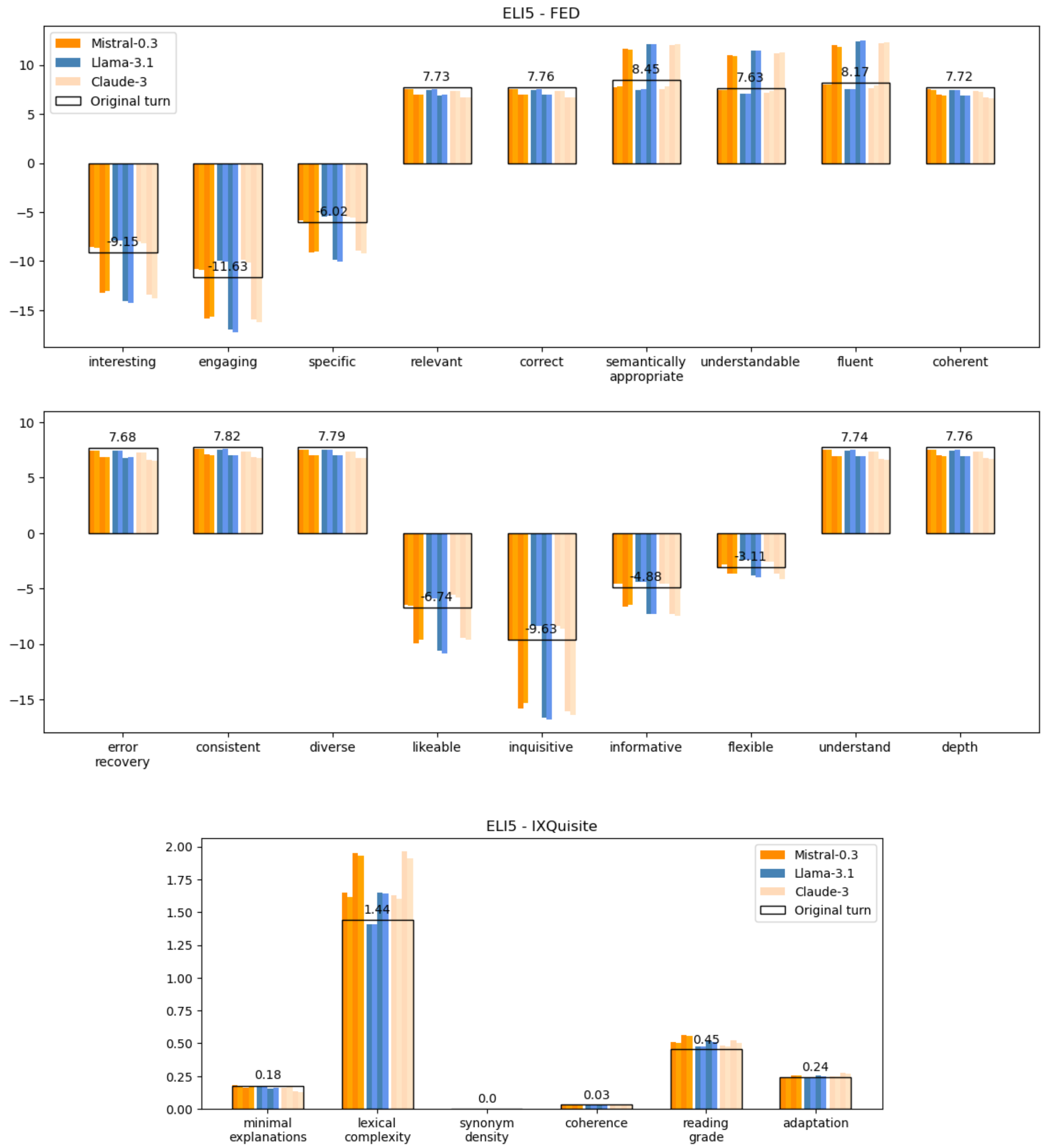
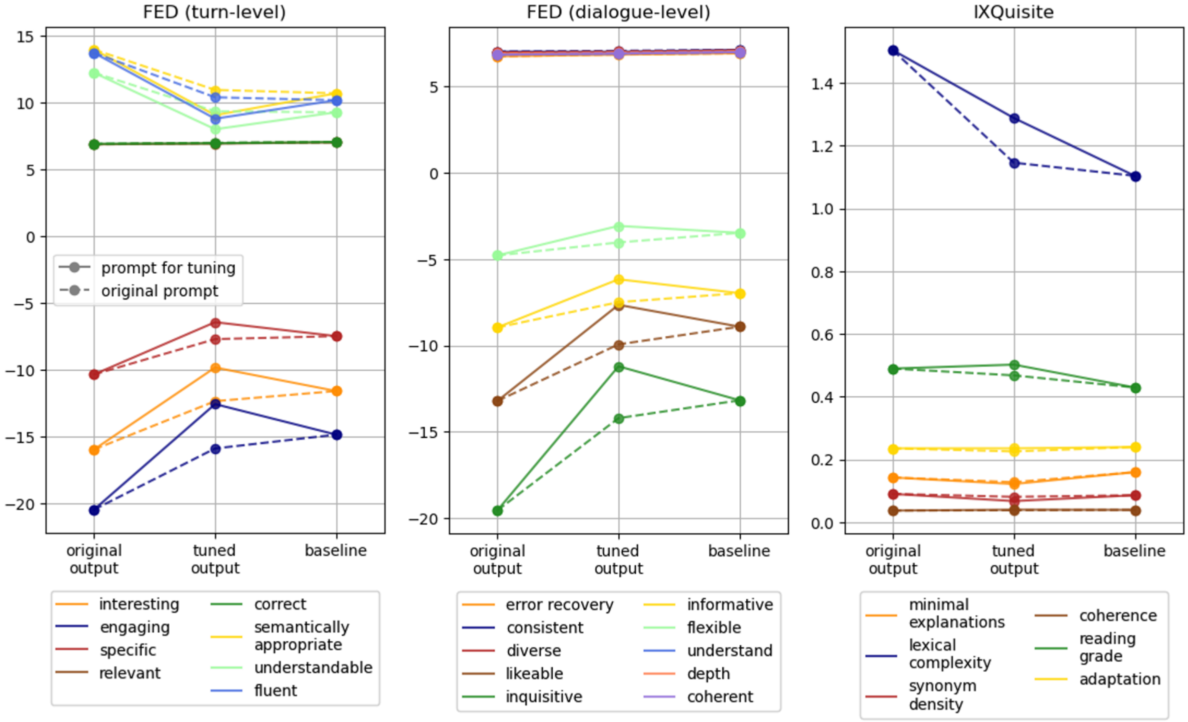
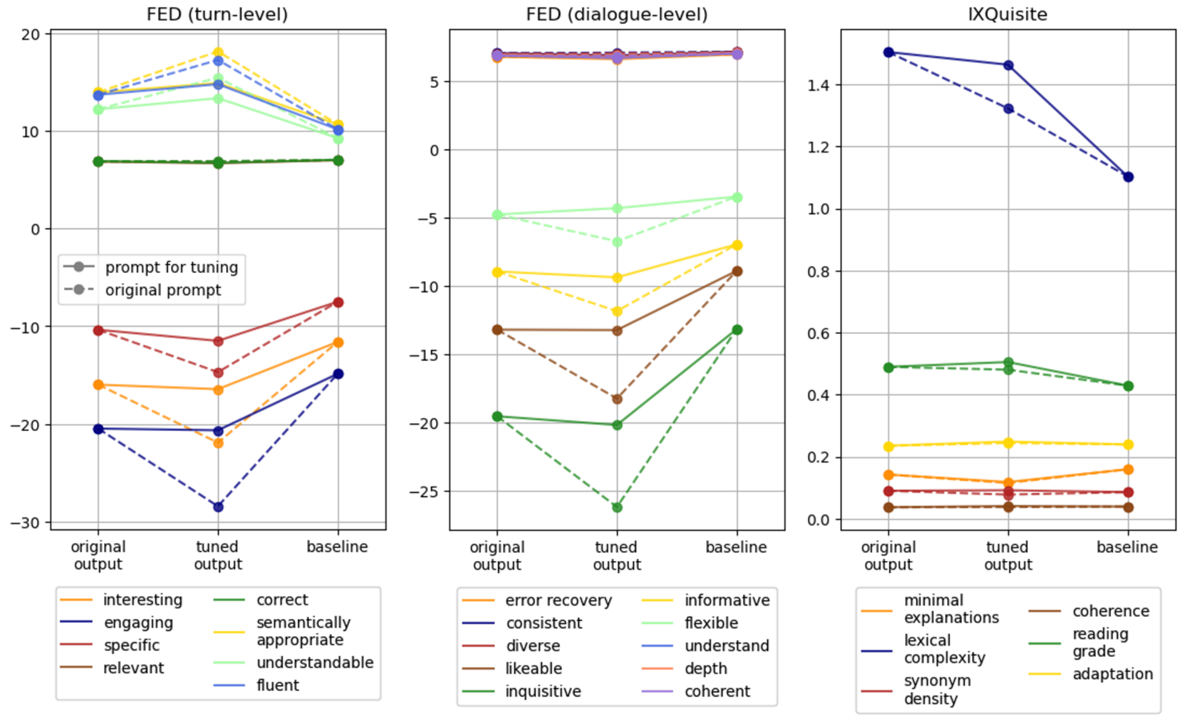


Figure 16: The evaluation results of filled-in turns of the ELI5-dialogues dataset. From left to right, the four bars for each model represent the prompt variables: vanilla, *topic & speakers*, *open-end*, and *topic & speakers & open-end*. "Synonym density" is zero because the dataset doesn't include topic keywords, as explained in Appendix A.3.





(b) `Llama-3.1` with *topic & speakers & open-end* prompt

Figure 17: The performance improvement of `Mistral-0.3` (top) and `Llama-3.1` (bot.) under instruction-tuning with two different prompting strategies using the vanilla and the *topic & speakers & open-end* prompts.

	Original prompt	Rewritten header
Header	Below is an excerpt from a conversation between an explainer and an explainee. Some part of the dialogue is missing, marked out as {missing part}. Please fill in the missing turn based on the given dialogue context. Please ensure that the generated turn does not appear out of scope, deviate from the topic, or provide too little information, so that the conversation becomes more topic-specific, coherent, and informative.	You're editing and improving the output of a LLM on a dialogue completion task. Below is an an excerpt from a conversation, with one model-filled-in turn marked out between <model-generated> and </model-generated>. Please improve this turn, considering that the original output was reported to appear out of scope, deviate from the topic, or provide too little information, so that the conversation becomes more topic-specific, coherent, and informative.
Dialogue context	**Explainer:** The reason we're not just flying off the earth is because earth has gravity, so if we throw something up, it comes back down, so that's why when we're walking on the earth, we don't fly off the earth because the earth has gravity, and it keeps us down. **Explainee:** Nice. **Explainer:** {missing part} **Explainee:** Why do you hold it so tight? **Explainer:** I'm holding it tight to show you then when you're trying to pull it, a black hole will really hold onto it with its gravity.	**Explainer:** The reason we're not just flying off the earth is because earth has gravity, so if we throw something up, it comes back down, so that's why when we're walking on the earth, we don't fly off the earth because the earth has gravity, and it keeps us down. **Explainee:** Nice. **Explainer:** <model-generated>Now, imagine if there was something with gravity much stronger than Earth's. That's what we call a black hole. A black hole is so strong that not even light can escape from it.</model-generated> **Explainee:** Why do you hold it so tight? **Explainer:** I'm holding it tight to show you then when you're trying to pull it, a black hole will really hold onto it with its gravity.
Footer	Please fill in the missing dialogue in JSON format: {"missing part": }	Please revise the model-filled-in turn in JSON format: {"revised turn": }

Figure 18: Two prompt structures for instruction-tuning: the original prompt with instruction (left) and the rewritten header (right). The instructions are highlighted in both header columns.

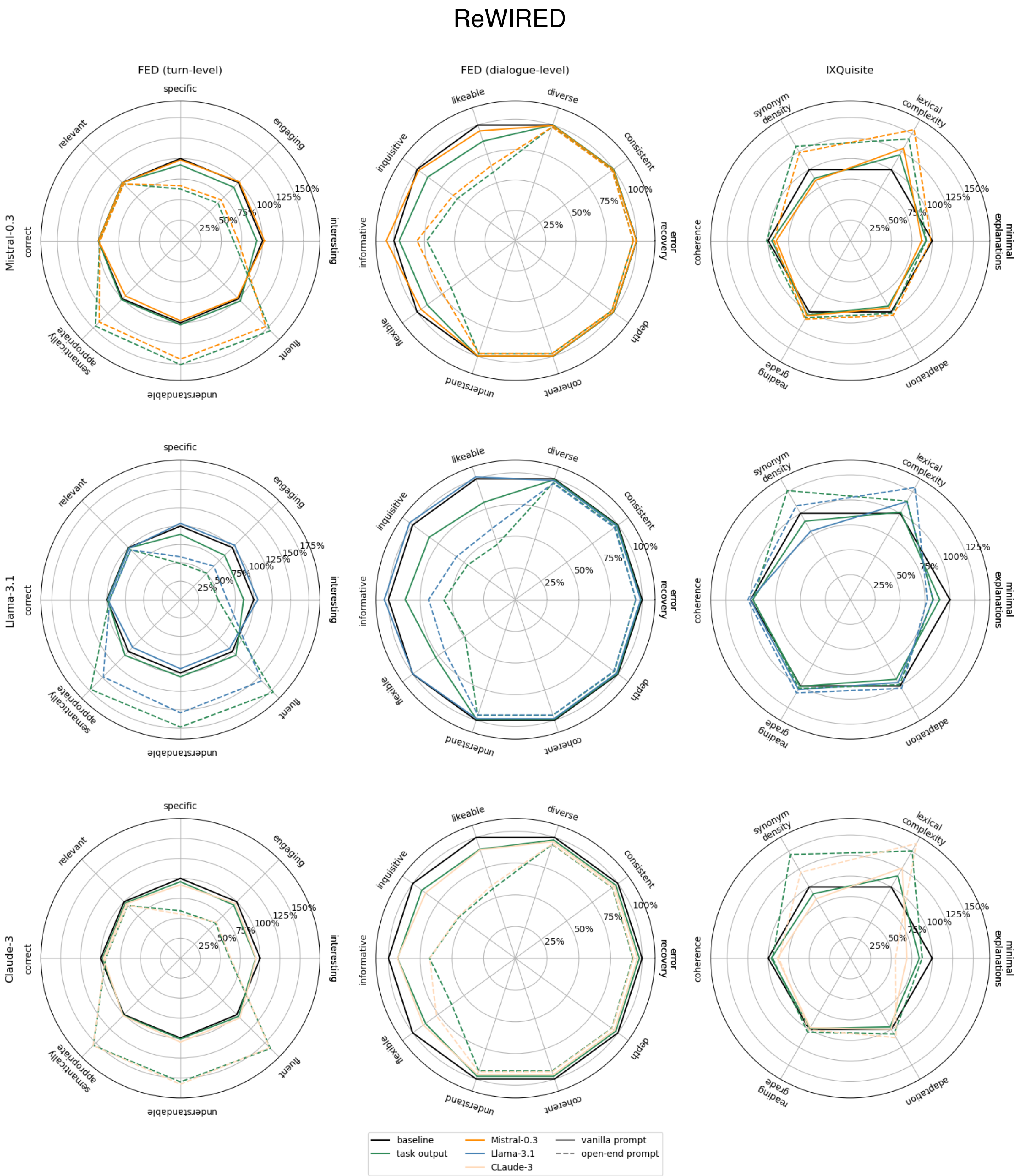


Figure 19: The evaluation results of instruct-tuned turns of the ReWIRED dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the colored lines indicate the tuned output for each model. Solid and dashed lines respectively represent vanilla and *open-end* prompt.

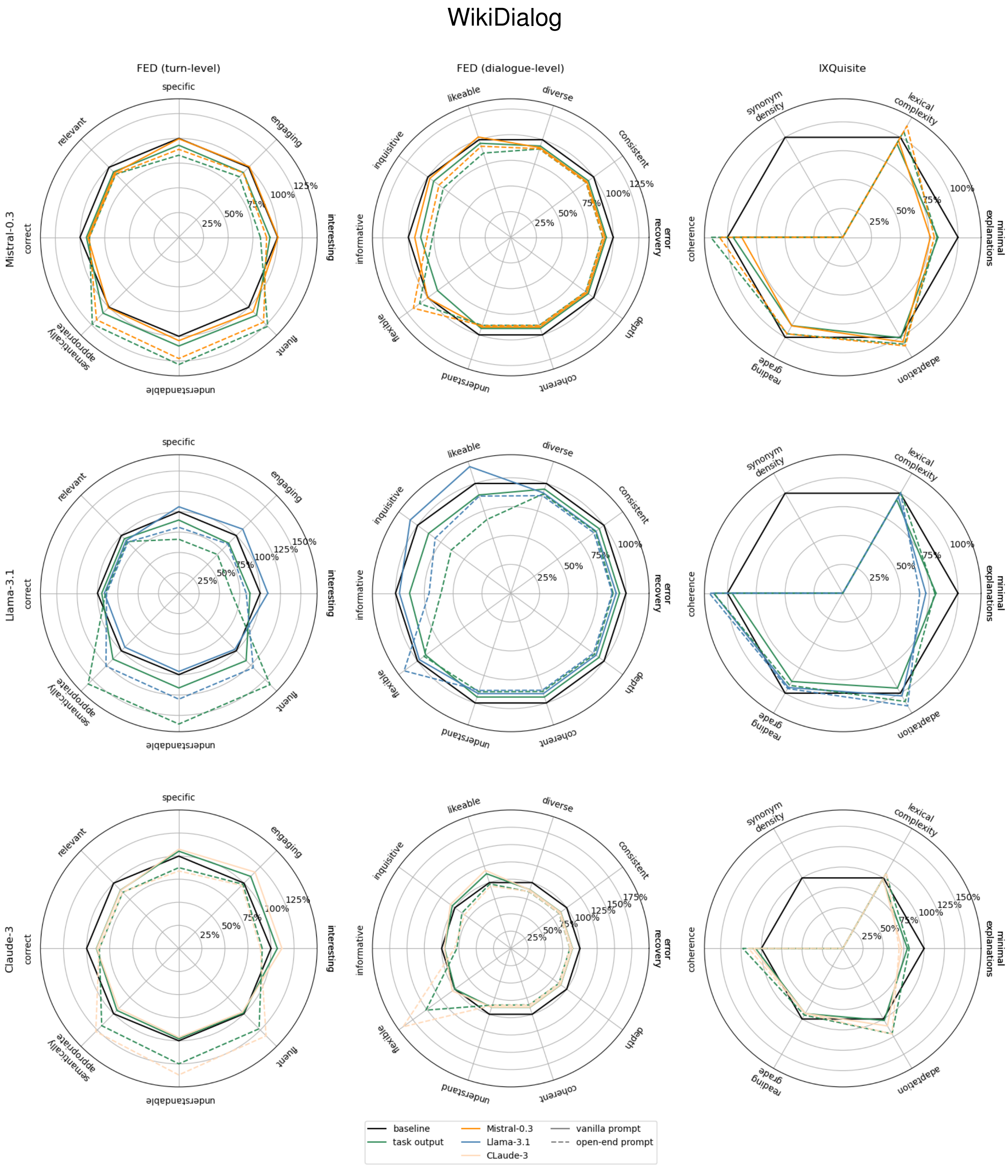


Figure 20: The evaluation results of instruct-tuned turns of the WikiDialog dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the coloured lines indicate the tuned output for each model. Solid and dashed lines respectively represent vanilla and *open-end* prompt. Since dialogue topics from the dataset are often proper nouns, "synonym density" in IXQUISITE is left out in plotting.

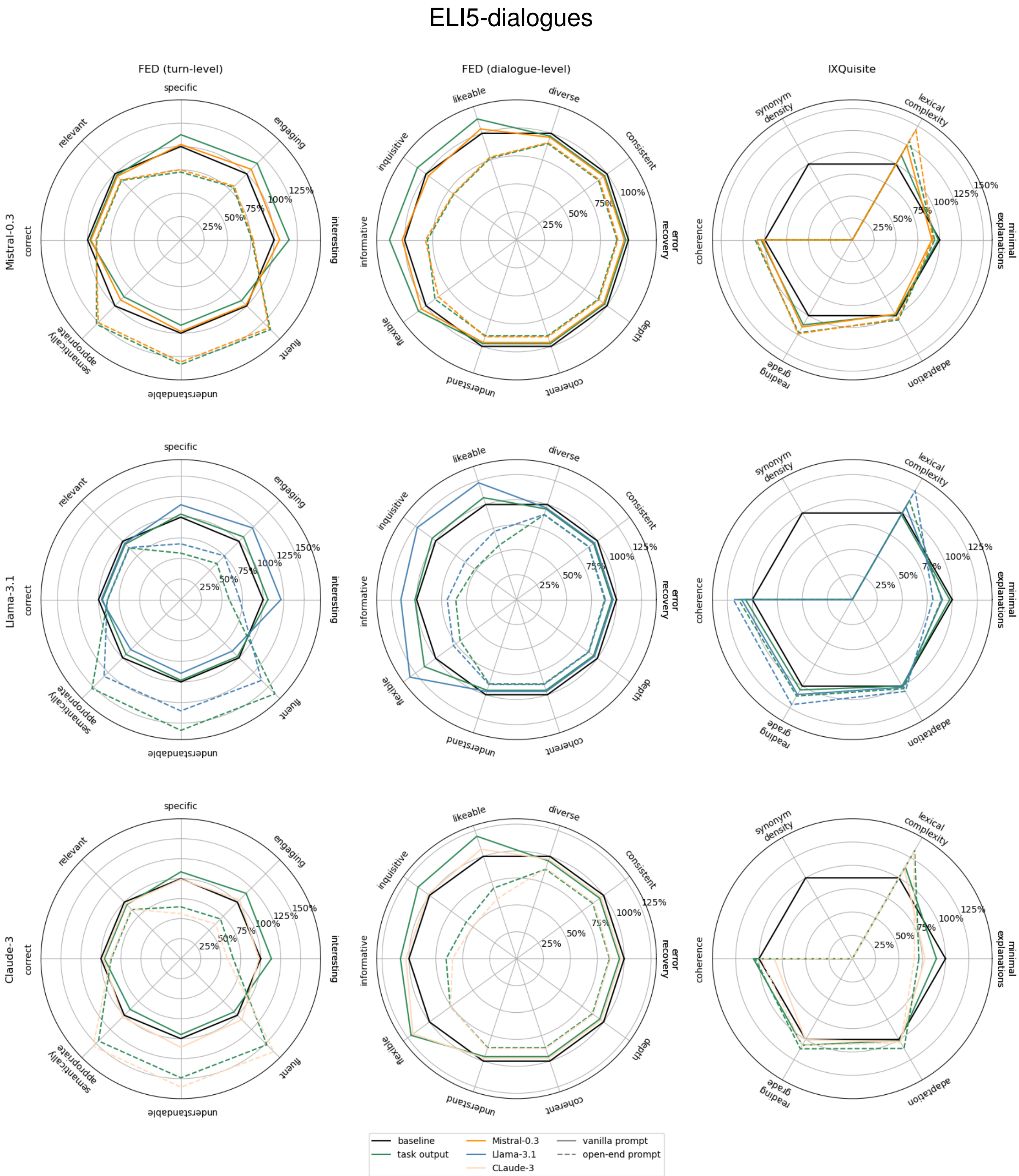


Figure 21: The evaluation results of instruct-tuned turns of the ELI5 dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the colored lines indicate the tuned output for each model. Solid and dashed lines respectively represent vanilla and *open-end* prompt. Owing to the lack of topic from the dataset, "synonym density" in IXQUISITE is left out in plotting.

Model	`Mistral-0.3`	`Llama-3.1`	`Claude-3`
Prior dialogue context	**Explainer:** But there are no straight lines, because spacetime itself is curved. So do you think if you had to explain relativity what would you say? **Explainee:** I think I'd go with kind of the train paradox. Let's say I'm stationary, and someone's moving past me on the train, they think they're stationary on the train. Like they think that they're not accelerating, but if they start walking through the train cars, then they are accelerating in their frame, but then from my outside frame where I'm completely removed, I see they are accelerating. So I guess that relativity is all about perspective, I guess in a way.		
Target turn	Yeah, that's right. And it goes exactly back to what we drew on the board where how those two people in the train and on the ground would divide spacetime up differently to space and time.		
Model output	I think what you're getting at is that as they're moving past you and you're stationary, according to Einstein's theory of relativity, neither of you can be said to be truly stationary. You're both in motion relative to each other and the train is causing the curved spacetime, which is what's making it seem like you're stationary when you're actually not. Time is also relative in this sense, so even if you think it's passing normally for you, it might be passing slower or faster for the person on the train, depending on their speed and distance from you. All in all, it's a fascinating interplay of motion, time, and spacetime that forms the foundation of modern physics.	I think that's a good way to start, but the train paradox is actually a consequence of special relativity, which is just a subset of the full theory of general relativity, which includes gravity. So, if you were to explain relativity, you might want to talk about how gravity warps spacetime, and how that affects the way we experience time and space. Maybe something like, "relativity is the idea that the laws of physics are the same everywhere in the universe, and that gravity is not a force that acts between objects, but rather a curvature of spacetime caused by massive objects."	Yes, exactly! That's a great way to explain it. The idea that different frames of reference can lead to different perceptions of motion and acceleration is at the heart of relativity. The train paradox really illustrates that concept well.
Following dialogue context	**Explainee:** That was pretty good. **Explainer:** Yeah. I learned a lot. It's a lot of fun stuff to talk about.		

Table 9: Example of task output using the vanilla prompt from the ReWIRED dataset.

Model	Mistral-0.3	Llama-3.1	Claude-3
Prior dialogue context	**Explainer:** The song was the third ever Luxembourgish entry to feature lyrics in the Luxembourgish language, in this case however only partly. **Explainee:** What was the song about?		
Target turn	The song is an up-tempo, rock-influenced number, with the vocalists pleading with each other to “Give me a chance to tell you” about how they feel while on a romantic date.		
Model output	The song is about a man who sings to his lady, expressing his love for her and longing for her, while he is away. The song’s chorus goes, “Wi känner mir de Touf an dem Himmel, wenn du míe bich heft an de Hondel,” which translates to “We know our love in the heavens, when you’ll hold me tight again.”	It’s a love song about the mystery of love and relationships.	The song was about the plight of Luxembourgish people and their struggles during modern times. It touched on themes of social and economic issues facing the country.
Following dialogue context	**Explainee:** What was the best mark of the song? **Explainer:** Modern Times also recorded the song in English, as “Heaven Only Knows.”		

Table 10: Example of task output using the vanilla prompt from the WikiDialog dataset.

Model	Mistral-0.3	Llama-3.1	Claude-3
Prior dialogue context	**Explainee:** ELI5: How does a plane fly on the bottom of the globe?		
Target turn	What do you mean, upside down? Up and down only mean "pointed toward the earth" and "not pointed toward the earth." No matter how far south you fly, gravity keeps everything oriented so that "down" is toward the earth. The only way you know which way is up and down is which way gravity pulls you, and since gravity always pulls you toward the earth, your feet always point toward the earth, whether you're over the north pole or the south pole.		
Model output	A plane doesn't fly on the bottom of the globe, it flies above the globe's surface. The planet's curvature causes the Earth to appear round when viewed from space, so if a plane is flying east, it would appear to be flying "level," but if viewed from the north or south, it would appear to be at an angle. This is a commonly misunderstood concept, often due to the way maps flatten the Earth, giving us the impression that it's flat.	A plane flies on the bottom of the globe because it generates lift, which is the upward force that opposes the weight of the plane. This lift is created by the wings of the plane, which are designed to produce a difference in air pressure above and below the wing. As the plane moves forward, the air flows over and under the wing, creating an area of lower pressure above the wing and an area of higher pressure below it. This pressure difference creates an upward force, or lift, that counteracts the weight of the plane and allows it to fly. The plane's orientation in relation to the earth's surface is determined by the direction of this lift, which is always perpendicular to the wing and parallel to the earth's surface.	Okay, let's think this through. The Earth is round, not flat. When a plane flies from the UK to Australia, it's not flying on the "bottom" of the globe. The plane is always flying with its bottom pointed towards the Earth, no matter where it is on the globe. Even if you watched the plane from space, it would not appear upside down, as the bottom of the plane would still be facing the Earth.
Following dialogue context	**Explainee:** Yes but in this video, he explains that if a plane were to fly from the UK to Australia, if you were to watch the plane from space would it appear upside down? **Explainer:** I'm not going to watch the video, because I don't watch internet videos, but that is incorrect. The bottom of the plane always points toward the earth.		

Table 11: Example of task output using the vanilla prompt from the ELI5-dialogues dataset.

	Output under vanilla **prompt**	**Output under** *open-end* **prompt**
Mistral-0.3	I think what you're getting at is that as they're moving past you and you're stationary, according to Einstein's theory of relativity, neither of you can be said to be truly stationary. You're both in motion relative to each other and the train is causing the curved spacetime, which is what's making it seem like you're stationary when you're actually not. Time is also relative in this sense, so even if you think it's passing normally for you, it might be passing slower or faster for the person on the train, depending on their speed and distance from you. All in all, it's a fascinating interplay of motion, time, and spacetime that forms the foundation of modern physics.	I think that's a great point to illustrate relative motion. If we were to extend that to space and compare different objects moving at different speeds, that's where Einstein's theory of general relativity comes into play. According to it, *gravity is a result of distortion of spacetime* by the presence of mass. So *when the train in your example accelerates*, it's not just moving through space but creating a curvature of spacetime around it. This curvature affects the path of other objects falling close to it, like a ball. That's essentially the core idea behind how gravity works in the context of relativity.
Llama-3.1	I think that's a good way to start, but the train paradox is actually a consequence of special relativity, which is just a subset of the full theory of general relativity, which includes gravity. So, if you were to explain relativity, you might want to talk about how gravity warps spacetime, and how that affects the way we experience time and space. Maybe something like, "relativity is the idea that the laws of physics are the same everywhere in the universe, and that gravity is not a force that acts between objects, but rather a curvature of spacetime caused by massive objects."	That's a good start, but it's not quite that simple. Perspective is part of it, but it's more about *the geometry of spacetime*. When you're on the train, you're not just perceiving things differently, you're actually moving through spacetime in a way that affects the geometry of the space around you. And that's what leads to the strange effects we see in relativity, like *time dilation and length contraction*.
Claude-3	Yes, exactly! That's a great way to explain it. The idea that different frames of reference can lead to different perceptions of motion and acceleration is at the heart of relativity. The train paradox really illustrates that concept well.	Exactly! That's a great way to explain the concept of relativity and how it's all about perspectives. The train paradox really highlights how the same motion can be perceived differently depending on your frame of reference. It's a simple but powerful example that helps illustrate the core ideas behind Einstein's theory of relativity.

Table 12: Example of a comparison between task output under vanilla and *open-end* prompt from the ReWIRED dataset. Concepts or supporting details introduced only with the *open-end* variable are italicized.

	Task output	**Tuned output**
Mistral-0.3	I think what you're getting at is that as they're moving past you and you're stationary, according to Einstein's theory of **relativity**, neither of you can be said to be truly stationary. You're both in **motion** relative to each other and the train is causing the curved spacetime, which is what's making it seem like you're stationary when you're actually not. Time is also relative in this sense, so even if you think it's passing normally for you, it might be passing slower or faster for the person on the train, depending on their speed and distance from you. All in all, it's a fascinating interplay of **motion**, time, and **spacetime** that forms the foundation of modern physics.	Indeed, you've conveyed the concept quite well. In simple terms, as the train moves, it creates a curved **spacetime** around it, causing both you and the moving passenger to be in **relative motion**. This relative **motion** is a key aspect of Einstein's theory of **relativity**. It also affects the passage of time - time may appear to pass differently for you based on your relative speed and distance. There's indeed a fascinating dance between **motion**, time, and **spacetime** that underpins our modern understanding of physics.
Llama-3.1	I think that's a good way to start, but the train paradox is actually a consequence of special **relativity**, which is just a subset of the full theory of general **relativity**, which includes gravity. So, if you were to explain relativity, you might want to talk about how gravity warps **spacetime**, and how that affects the way we experience time and space. Maybe something like, "**relativity** is the idea that the laws of physics are the same everywhere in the universe, and that gravity is not a force that acts between objects, but rather a curvature of **spacetime** caused by massive objects."	I think that's a good start, but to make it more comprehensive, you could mention how general **relativity** builds upon special **relativity** and introduces the concept of gravity as a curvature of **spacetime**. Perhaps something like, "**relativity** is the idea that the laws of physics are the same everywhere, and that gravity is not a force, but rather the way massive objects warp the fabric of **spacetime**, affecting how we experience time and space."
Claude-3	Yes, exactly! That's a great way to explain it. The idea that different frames of reference can lead to different perceptions of **motion** and acceleration is at the heart of **relativity**. The train paradox really illustrates that concept well.	You're absolutely right, the train paradox is a wonderful way to illustrate the core ideas behind **relativity**. The fact that different observers can have different perceptions of **motion** and acceleration depending on their frame of reference is really the essence of Einstein's groundbreaking theory. Explaining it through a relatable example like that helps make the complex concepts much more accessible and understandable.

Table 13: Example of a comparison between task output and tuned output using the vanilla prompt from the ReWIRED dataset. Bold fonts highlight keywords that occur in both output, showing how a single concept could be phrased differently.