

Evaluating LLM-generated Explanatory Dialogue Turns through Dialogue Completion

Yi-Sheng Hsu^{*,†} Nils Feldhus^{2,3,*,‡} Manfred Stede¹

¹Applied Computational Linguistics, University of Potsdam

²CLCG, University of Groningen ³German Research Center for Artificial Intelligence (DFKI)
n.feldhus@rug.nl stede@uni-potsdam.de

Abstract

Synthetic explanatory dialogues hold promise for applications from tutoring systems to model rationalization, yet their quality is often deemed inferior to human-authored conversations. This study focuses on dialogue completion, where LLMs are instructed to complete a missing dialogue turn within a given context of an explanatory conversation. Using the task as a lens to investigate the nuanced differences between human-spoken, human-written, and synthetic explanatory dialogue. By evaluating LLM-generated turns against original turns using the FED and IXQuisite test suites, we uncover a significant bias in the FED metric towards written language, which may penalize the interactivity of spoken dialogue. This pattern is replicated by three popular LLMs (Mistral, Llama, Claude), which excel in fluency and yet struggle with conversational engagement, particularly on completing spoken dialogue transcripts. More importantly, we demonstrate that these models can significantly improve their performance through a zero-shot, iterative self-refinement method. By providing feedback based on their specific weaknesses, we show that LLMs can iteratively adapt their outputs to better resemble the interactive qualities of human dialogue.

1 Introduction

Explanation is a critical component of conceptual understanding (Miller, 2017) and the generative capabilities of large language models (LLMs) have created new opportunities for interactive explanatory dialogues in domains like tutoring and instruction (Xu et al., 2024; Li et al., 2024), but modeling the dynamic, back-and-forth flow of conversation remains a significant challenge (Ni et al., 2023).

Focusing on explanatory contexts, our study explores the dialogue completion task (Figure 1) and

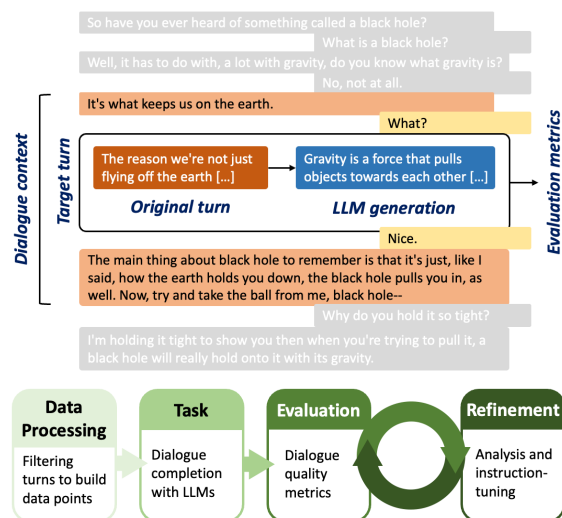


Figure 1: The summary of the dialogue completion task (top) and our workflow (bottom). In the task, we extract the surrounding context of a target turn, prompt an LLM to fill in the removed turn, and evaluate the original and generated turn with test suites.

probes into how models and humans act differently in a dialogue, as well as the measurement of such differences (Dai et al., 2022; Li et al., 2023; Wang et al., 2024). Building on these motifs, we investigate LLMs’ capability of performing explanation on given topics in conversational scenarios, contributing the following:

- (1) We conduct a comparative analysis of explanatory dialogues across three modalities – human-spoken, human-written, and synthetic – using LLM-based dialogue completion as a probe, revealing a previously undiscussed bias in the popular FED metric (Mehri and Eskénazi, 2020a), which favors written text and tends to penalize transcribed spoken conversations (§3.3).
- (2) We systematically evaluate three representative LLMs, finding that their strengths (fluency, coherence) and weaknesses (engagement, specificity) vary significantly depending

*Shared first authorship.

†Work done at University of Potsdam.

‡Work done at Technische Universität Berlin.

on the dialogue modality they are tasked to complete. This highlights the challenge models face in capturing the pragmatic nuances of different conversational contexts. (§3.2).

- (3) We validate a lightweight self-refinement method where LLMs refine their own generated turns based on automated, metric-driven feedback. This demonstrates that models can effectively self-correct their initial weaknesses in interactivity, narrowing the performance gap with human-written dialogues (§3.4).

Methodologically, we suggest that dialogue completion is a useful probe for both LLM behavior and metric behavior. The approach benefits further studies on explanatory dialogues, interactivity, metric design, and multi-modal evaluation.¹

2 Background

2.1 Explanatory dialogue

Real-life scenario. Explanatory dialogues typically exhibit linguistic features distinct from plain text at syntactic, semantic, and pragmatic level owing to the interaction between speakers. 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, being mindful of the expertise level and familiarity with a topic becomes essential (Wachsmuth and Alshomary, 2022). Furthermore, 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.

Synthetic dialogues. Prior work suggested that human dialogues were generally seen as more applicable than synthetic ones (Dai et al., 2022; Li et al., 2023; Stacey et al., 2024). Although model-generated explanatory dialogues were later found more favorable than human ones, they were best viewed as augmenting rather than replacing human expertise (Li et al., 2024). Similarly, it was noted that model-generated dialogues may already benefit interactive systems in fine-tuning models towards producing more natural dialogues (Dai et al., 2022) under common conversational topics (Chen et al., 2023). Concerning explanatory dialogues, LLMs

were also found to benefit educational scenario Tack and Piech (2022), where LLM-based assistance was provided to both teacher (Maurya et al., 2025; Carpenter et al., 2024) and student roles (Liu et al., 2024) in various classroom settings.

2.2 Evaluation of dialogues

Measuring dialogue quality is challenging owing to the interactive nature involving speakers and turns, covering criteria such as coherence, participation, and engagement (Adiwardana et al., 2020). Conventional methods for evaluating NLG output could hardly capture the quality of conversational flow as how human would perceive it (Deriu et al., 2021). The failure of conventional reference-based metrics (F_1 , BLEU) directed prior work to develop approaches that remain robust under the dynamic conversational expressions (Zhang et al., 2021; Ma et al., 2022; Mendonça et al., 2024). From a didactics-oriented perspective, Anagnostopoulou et al. (2025) proposed the IXQUISITE metric suite for explanatory dialogues, employing count-based methods to track linguistic features.

Model-based, reference-free metrics have become a popular alternative. Both USR (Mehri and Eskénazi, 2020b) and FED (Mehri and Eskénazi, 2020a) leverage pretrained dialogue models like DialoGPT to score turn and dialogue-level quality. While alternatives like DYNAEVAL (Zhang et al., 2021) and SELF-EVAL (Ma et al., 2022) exist, they have not shown a significant advantage over FED for overall quality assessment, motivating our use of FED for pragmatic evaluation alongside IXQUISITE’s linguistic features.

2.3 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 for rationalizing models’ decision-making. Chain-of-thought (Wei et al., 2022) involved 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.

3 Experiments

We instruct LLMs to complete explanatory dialogues. Figure 1 (bottom) presents an overview

¹Our code and data is available at: https://github.com/doyouwantsometea/dialog_completion

of our workflow, which consists of four major stages: Data processing (Appendix A.1), Dialogue completion task (§3.1 & §3.2), Evaluation (§3.3), and Iterative self-refinement (§3.4). The experiments cover three English-language datasets featuring two-agent explanatory dialogues: ReWIRED (Anagnostopoulou et al., 2025), 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 Dialogue Completion Task

We first filter a dialogue D for all explainer turns longer than a minimum length l . For each such target turn T_i , we construct a dialogue context d_i consisting of the w preceding and w succeeding turns, replacing T_i with a placeholder. An LLM is then prompted to fill in this missing turn, producing a completed dialogue context d'_i . Figure 8 in the Appendix visualizes the process and further demonstrates how data instances are built out of the original (*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; 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 human conversations. 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 in the main experiments, with an additional ablation study exploring $w = 4$ (§ 5.1). A window of $w = 2$ balances context and input length. It captures the most recent full exchange between the explaineer and explainer, providing immediate conversational context while avoiding less relevant distant turns and potential model context limits associated with larger windows. Following a rough analysis on the output from a trial run (Appendix C), l is set to 30 tokens.

Operational definition. We define an *explanatory turn* as any explainer-role turn of at least l tokens in a dialogue whose corpus-level purpose is explanation, which holds for all three datasets (Appendix A.1). Explanation acts (concept, reason,

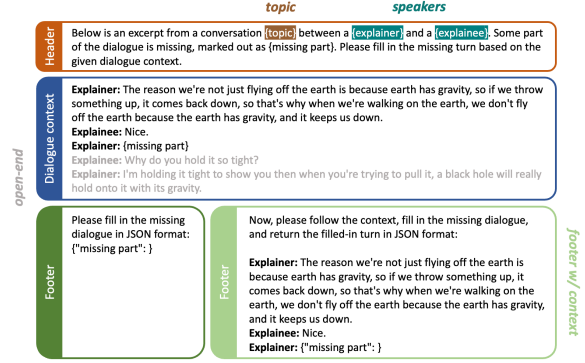


Figure 2: Prompt structure for the explanation dialogue completion task with all the tested prompt variables marked out. The example dialogue is from ReWIRED.

evidence) are considered intertwined (Alshomary et al., 2024) and not separately annotated. Small talk, clarifications, and comprehension checks naturally co-occur with explanation, and we keep such turns intentionally so that the task reflects realistic dialogue complexity. In a manual spot-check of 30 random ReWIRED target turns, 27 were strictly explanation-bearing and all 30 at least partially so; the partial cases were explainer comprehension checks, a natural component of extended explanations.

3.2 Prompting

We conduct experiments with Mistral-0.3 7B (Jiang et al., 2023) and Llama-3.1 8B (Grattafiori et al., 2024) on an 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.

As shown in Figure 2, we explore a few variables in structuring the prompt. In the experiments, *topic* & *speakers* and *open-end* are used for adapting the prompt from its vanilla setting. The detailed descriptions of the prompt variables and their combinations are described in Appendix B and C.

3.3 Evaluation

We apply FED (Mehri and Eskénazi, 2020a) and IXQUISITE (Anagnostopoulou et al., 2025) (§2.2) 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 the aggregated 24 aspects.

FED assesses explanatory dialogues with multiple positive and negative feedback utterances that

could reflect the perceived quality for 18 aspects²: 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 (Anagnostopoulou et al., 2025) 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

3.4 Iterative self-refinement

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 $\frac{F_m + F_o}{n_{(m+o)}}$, i.e., including both scores of model output and original turns.

For each instance, we 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 descriptions⁴ to form instructions. In addition, two **strategies** are tested for the overall prompt structure (Figure 9): (1) Modifying the original prompt (Figure 2) by adding the instruction to the end of the header; (2) Rewriting the header such that the model has to revise

²Table 5 in the Appendix lists the 18 aspects covered by the FED metric, categorized into turn- and dialogue-level.

³Details are provided in Table 7 in the Appendix.

⁴Table 12 (App.) shows the descriptions for all 24 aspects.

	ReWIRED	WikiDialog	ELI5-dialogues
Built data points ($l=30, w=2$)	743	2,613	2,407
Mistral-0.3	398 (53.6 %)	1,941 (74.3 %)	1,503 (62.4 %)
Llama-3.1	521 (70.1 %)	2,149 (82.2 %)	1,675 (69.6 %)
Claude-3	524 (70.5 %)	2,053 (78.6 %)	916 (38.1 %)

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

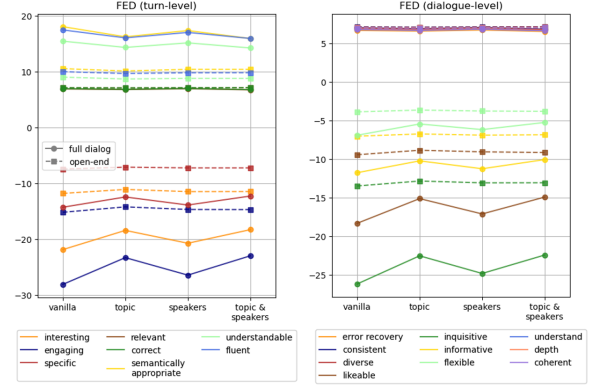


Figure 3: Dialogue quality measured by FED of Llama-3.1 across *topic*, *speakers*, and *open-end* variables in prompt design.

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).

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 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.⁵ Enforc-

⁵For each dataset (ReWIRED, WikiDialog, and ELI5-dialogues), a sample instance and corresponding model output is provided in Appendix A.1.

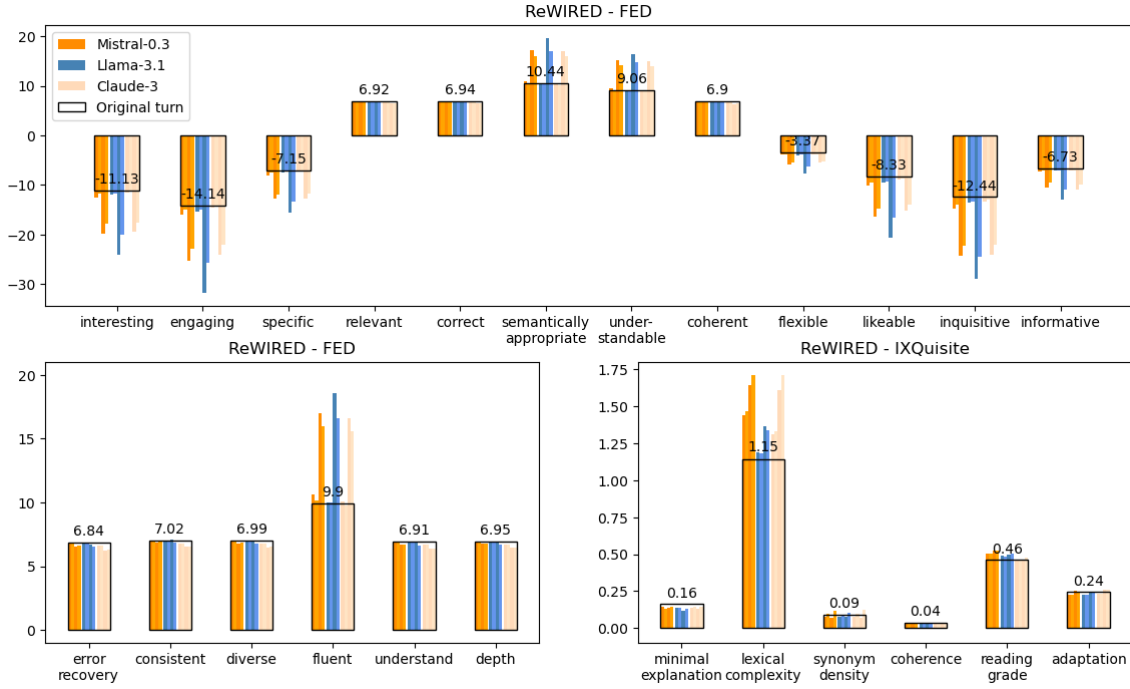


Figure 4: Evaluation of the LLM-generated and the original human turns with FED (top and bottom-left) and IXQUISITE (bottom-right) on ReWIRED. The four bars of each model represents a prompt variant: *vanilla*, *topic & speakers*, *open-end*, and *topic & speakers & open-end*. The black box denotes baseline, i.e., the original turns.

ing a JSON output facilitates parsing but can cause task failures if a model’s output exceeds the token limit before closing the object. As seen in Table 1, this issue is most prominent for Claude-3 on the ELI5 dataset, likely because its longer, written-style turns induced stronger truncation.⁶

Task performance. We compare the metric scores of LLM-generated and original turns, discovering that LLMs tend to formulate explanatory turns well while interacting with the interlocutor rather poorly. Such trend is universal across models and datasets, but the gap between LLM-generated turns and baseline may vary. For instance, the gap is the widest for ReWIRED (Figure 4), while the differentiation of WikiDialog and ELI5 appear closer (Appendix D). The three models exhibit inconsistent performance: Turns generated by Claude-3 are often rated similarly to the baseline and yet sometimes outperformed by those produced by the two open LLMs. Llama-3.1 slightly outperforms Mistral-0.3 in most FED aspects, which is consistent with the trial run (Appendix C). Furthermore, 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 (Figure 4). In contrast, *topic & speakers* perturbs the output less, and the relative value compared to the baseline fluctuates (Figure 3).

To confirm the reliability of these trends, we conducted paired t-tests and calculated Cohen’s d on the ELI5 dataset ($N = 854$). We find a statistically significant trade-off ($p < 0.001$) between fluency and interactivity. For instance, Llama-3.1 significantly dominates on surface-level metrics like Fluency ($p < 10^{-4}$, $d = 0.13$) and Understandability ($p < 10^{-5}$, $d = 0.15$). Conversely, Mistral-0.3 dominates on conversational metrics like Inquisitiveness ($p < 10^{-3}$, $d = 0.13$) and Informativeness ($p < 10^{-6}$, $d = 0.17$). This statistically confirms that models with high surface fluency can simultaneously fail at pragmatic engagement.

4.2 Refining filled-in turns

Table 2 presents the worst-performing features across different datasets and models, highlighting that filled-in dialogue turns tend to perform poorly in areas such as specificity, flexibility, and consistency.⁷ While comparing the two prompting

⁶Task accomplishment reflects strict formatting adherence, while the capability of the models to generate valid explanations remains high even in these truncated cases.

⁷Although IXQUISITE scores do not directly imply dialogue quality, we do not exclude them in feature selection

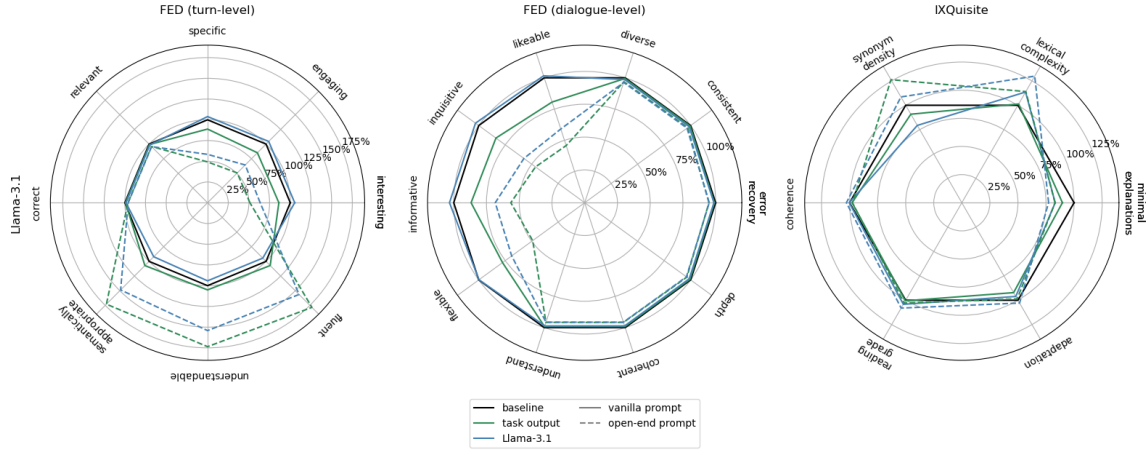


Figure 5: The evaluation results of refined turns of Llama-3.1 on the ReWIRED dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the blue lines indicate the refined output of Llama-3.1. Solid and dashed lines respectively represent vanilla and *open-end* prompt.

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

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.

strategies used to specify the identified worst features, we find that using a refining-specific prompt leads to more substantial improvements than simply adapting the original prompt (Appendix D). This approach also aligns better to the self-refine prompting framework (Madaan et al., 2023). On the other hand, with the partially eliminated context with prompt variable *open-end*, iterative self-refinement introduces more drastic performance changes but also uncertainty in that the self-refined output can be rated worse more often.

Despite occasional uncertainties, iterative self-refinement produces statistically significant improvements on most FED metrics, with the largest gains on conversational aspects such as inquisitiveness and engagement, while fluency declines – a trade-off indicating that refinement shifts the model’s style from a static, textbook-like format towards a more interactive dialogue. Table 3 confirms this pattern across all nine dataset $\times$ model because the FED aspects remain more dominant.

Dataset	Model	N	Eng. $\uparrow/7$	Flu. $\downarrow/3$	Oth. $\uparrow/8$	Oth. $\downarrow/8$
ReWIRED	Llama-3.1	521	7	3	0	3
	Mistral-0.3	293	7	3	0	0
	Claude-3	611	7	3	0	8
WikiDialog	Llama-3.1	2,044	7	3	0	8
	Mistral-0.3	1,572	7	3	0	8
	Claude-3	2,428	5	3	0	8
ELI5	Llama-3.1	1,708	7	3	8	0
	Mistral-0.3	1,213	6	3	8	0
	Claude-3	1,225	6	3	0	0

Table 3: Cluster-grouped refinement effects per dataset $\times$ model cell (vanilla prompt): the number of FED metrics per cluster with a significant change after iterative self-refinement (paired t-test, $p < 0.05$). Engagement (Eng.) = interesting, engaging, specific, likeable, flexible, informative, inquisitive; Fluency (Flu.) = semantically appropriate, understandable, fluent; Other (Oth.) = the remaining eight FED metrics (relevant, correct, coherent, error recovery, consistent, diverse, depth, understanding).

combinations on the vanilla prompt: engagement improves significantly and fluency drops in every combination. The remaining FED metrics move up on ELI5 but down or not at all on WikiDialog and ReWIRED; they follow dataset formality – precisely where FED’s written-text bias surfaces (§5.2) – whereas the engagement gains appear regardless of dataset, indicating that they come from the refinement itself.

We discover that the refined output tends to achieve higher scores in most FED aspects, but when LLMs outperform human explanations (e.g., “semantically appropriate,” “understandable,” “fluent”), scores may decrease. Figure 5 shows that Llama-3.1 follows such trend; although it usually performs worse on the task, iterative self-refinement substantially improves its output, even under scenarios where the initial output already

appears promising (Figure 20). In comparison, Claude-3 often achieves the best scores from the start, limiting its potential of further refinements. Such trend is also consistent across datasets: Substantial self-refinement is more effective when the gap is larger between task performance and baseline. With the same vanilla prompt, the output of ELI5 performs consistently pre and post-refinement in comparison to the other two (Appendix D).

5 Discussion

5.1 LLM output of dialogue completion

In the dialogue completion task, LLMs’ outputs align more closely with the provided dialogue context than with additional information appended to the task description, as omitting later turns increases output variability. Using *open-end* prompt variable consistently yields poor results, confirming the gap between explanations addressed by human speakers and LLMs, as well as highlighting how LLMs struggle with reduced context: Examples of *open-end* (Appendix C) 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 may still come up with explanatory turns that approximate the human ones within their output space.

To further investigate the impact of context size, we conducted an **ablation study** increasing the context window to $w = 4$ using Claude-3. Increasing the context significantly improved interactive metrics on spoken data, increasing Engagement by +6.56 points and Inquisitiveness by +7.06 points on ReWIRED. This suggests the passivity of LLMs is partly due to insufficient context. However, this gain came at a cost: $w = 4$ outputs suffered a significant drop in Fluency (-3.71) and Semantic Appropriateness (-3.93). Because doubling the input context doubles inference costs and degrades surface fluency, our proposed iterative self-refinement method offers a more token-efficient path to achieving similar interactivity gains (e.g., Inquisitive +5.86 on WikiDialog).

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 4, but most of the time, prompt variables *topic* and *speakers* hardly affect FED and IXQUISITE scores.

5.2 Dialogue quality of datasets and FED bias

This subsection uses human reference turns only – no model outputs are involved – keeping the metric-bias analysis separate from the refinement analysis, which §5.4 cross-validates with signals independent of FED. Our cross-modal analysis reveals a problematic characteristic of the FED metric: a strong preference for written dialogues over transcribed spoken ones. As shown in Appendix A.2, the human-written ELI5 corpus consistently outperforms the human-spoken ReWIRED corpus across nearly all FED aspects, including interactive ones like “engaging” and “likeable”. This is counter-intuitive, as spoken conversation is arguably the more natural and interactive medium, and suggests that FED, likely due to the nature of DialogPT’s training data, conflates textual formality and polish with conversational quality. While IXQUISITE confirms that written texts use more complex language, the universal superiority of written text in FED raises critical questions about its suitability for evaluating spoken dialogue systems, as it may unfairly penalize the less structured, more dynamic nature of human speech.

Surface-feature regression diagnostic. To quantify the bias, we correlate the FED scores of the human reference turns with four surface features of the same turns: turn length, mean word length, punctuation density, and type–token ratio (Table 8 in the Appendix). The engagement and fluency clusters move in opposite directions on every one of these features: engagement scores favor longer turns with more repeated words, while fluency scores favor shorter, lexically denser turns. FED scores thus partly depend on surface form alone, matching the hypothesized bias and offering a handle for mitigation, e.g., length-normalized FED scoring (cf. §5.4).

LLM-as-a-Judge validation. To validate this observed metric bias without relying solely on

Evaluated Aspect	Mistral-0.3	Llama-3.1	Claude-3
Relevant	-43.7	-38.8	-62.8
Interesting	-41.9	-57.8	-26.2
Fluent	47.9	60.4	37.7
Correct	-42.9	-37.7	-62.5
Understandable	46.3	59.1	35.9
Semantically appropriate	42.2	57.4	30.0
Specific	-44.7	-56.6	-28.9
Engaging	-38.5	-53.7	-15.5
Likeable	-38.1	-53.9	-16.3
Understand	-43.5	-38.6	-62.7
Coherent	-43.7	-38.8	-62.8
Depth	-43.2	-37.9	-62.6
Flexible	-8.7	-10.7	7.2
Diverse	-42.0	-36.4	-62.3
Inquisitive	-45.4	-57.9	-31.4
Error recovery	-44.9	-40.5	-63.0
Informative	-38.1	-44.4	-32.2
Consistent	-41.4	-35.6	-62.2
Minimal explanations	-19.2	-17.5	-25.8
Lexical complexity	28.9	2.2	19.1
Synonym density	1.2	0.0	0.9
Adaptation	2.2	-14.0	13.1
Reading grade	1.8	-11.9	-13.0
Coherence	3.3	2.3	4.8

Table 4: Normalized evaluation results in percentage (%) of each model on all datasets using the vanilla prompt. The best scores per aspect are highlighted. The normalized scores do not denote the relative deviation from the baselines, but instead the relative performance among all the measured aspects.

DialoGPT, we conducted an **LLM-as-a-Judge** study on a stratified sample of 300 turns (100 per dataset): a blind ranking of the human, task, and refined turns by Gemini 2.5 Pro, instructed to prioritize interactivity, clarity, and coherence over surface fluency (setup in Appendix E). We treat the judge as supplementary rather than primary evidence: LLM judges have biases of their own (positional, verbosity, self-preference), judging all experimental cells would be prohibitively expensive, and API models change over time whereas DialoGPT is fixed and reproducible. While FED frequently rated model outputs higher, the judge preferred the original human turn in 65.7% of cases on the spoken ReWIRED dataset (vs. 20.2% for model output). A second judge, Claude Haiku 4.5, preferred the human turn on ReWIRED even more often (85%), with fair agreement between the two judges (Appendix E). This convergence across two evaluator families supports our core claim: automated metrics exhibit a written-text bias that penalizes the natural interactivity of spoken dialogue.

5.3 Model-generated and human turns

LLMs generate functional explanations but often fail at interactivity, confirming the gap between synthetic and human dialogues (Dai et al., 2022; Stacey et al., 2024). Models perform better on dialogues that align with their typical output style, such as being grammatically correct and elaborate. Over-

all, 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; instead, the differentiation between WikiDialog and ELI5 imply that the detailed expressions (ELI5) may contribute to better 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”), matching the insufficient specificity of generated dialogue pointed out by Wang et al. (2024). Table 4 reveals that, across the evaluated aspects, the three models exhibit divergent performance. The normalized scores in Figure 4 reveal that, although seemingly better at first glance, Claude-3 does not significantly outperform the open LLMs, but it is rated higher at interaction (“engaging”, “likeable”), which is commonly considered challenging for dialogue systems. 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).

On the other hand, the IXQUISITE metric reveals that models exhibit distinctive linguistic patterns, e.g., Mistral-0.3 favors more complex expressions, while Claude-3 employs slightly more conjunctions. Alongside FED, we observe that better interactivity could correlate to the usage 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

Our results confirm that LLMs can self-refine their outputs via zero-shot prompting. The different evaluation scores for refined outputs, which often contain similar information to the original (Appendix A.1), underscore that effective explanation depends not just on what is said, but how it is phrased and integrated into the conversation (Wachsmuth and Alshomary, 2022). While Llama-3.1 consistently yields substantial improvements for the FED aspects (Figure 5), the other two models sometimes score worse after iterative self-refinement. This tendency could be related to the task performance: Larger gap between output and baselines leads to more room for improvement. Such an observation applies not only to

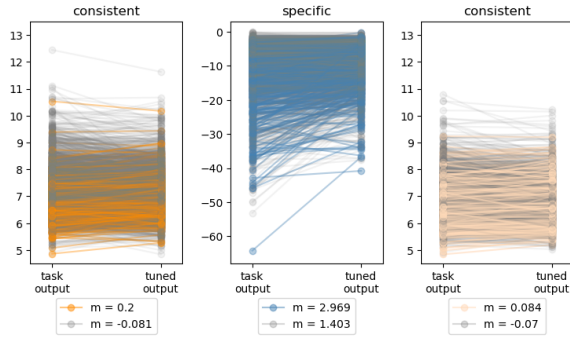


Figure 6: Instance-level score change after iterative self-refinement (l: Mistral-0.3, c: Llama-3.1, r: Claude-3), w.r.t. their worst-performing feature (Table 2) on ReWIRED. Changes in scores are represented by slope m . Model-specific colors highlight instances where the given aspect is described in the instruction.

cross-model comparisons but also datasets (Appendix D). Furthermore, instructions focusing on certain aspects in fact introduce an overall improvement, owing to reference utterances shared across multiple FED aspects; the same overlap shows in the sensitivity to prompt variables (Figure 5), where different FED aspects exhibit similar tendencies.

On instance level, specifying aspects in the prompt yields substantial score increases: Figure 6 shows a steeper gradient where the given aspect is described, and across all prompt-variable combinations on ReWIRED, the prompted instruction correlates with the quality on the corresponding evaluated aspect.

Crucially, the self-refinement process shows the most significant gains on the exact metrics where LLMs initially struggled (e.g., “specific”, “inquisitive”, “flexible”), as seen by comparing Figures 4 and 5. This demonstrates that the refinement is not a generic improvement but a targeted correction of identified deficiencies. The fact that scores sometimes decrease on metrics where LLMs were already “superhuman” (e.g., “fluent”) suggests the model is learning to trade some of its sterile perfection for more natural, human-like expression.

IXQuisite-only feedback ablation. Since FED both scores the outputs and partly supplies the refinement feedback, the gains could merely reflect optimizing towards FED. To test this, we re-run refinement using only the six IXQUISITE features for the worst-feature selection (no FED signal) on Claude-3 $\times$ ELI5 with the vanilla prompt, and re-evaluate the outputs with the full pipeline.⁸ With-

⁸The revising LLM in this ablation is Claude Haiku 4.5, since Claude-3 Haiku had meanwhile been deprecated; the

out FED feedback, the engagement improvements become smaller, as expected when a relevant signal is removed, but all seven engagement metrics still improve significantly, and the fluency trade-off remains as well (Table 10 in the Appendix): the engagement gains are not an artifact of FED-derived feedback, and the fluency trade-off is a property of the refinement task itself.

Bias-corrected refinement effects. To test whether the refinement gains are carried by the surface features behind FED’s written-text bias (§5.2), we re-score FED with the surface contribution removed: a linear regression fitted on human reference turns predicts each metric from the four surface features, and this prediction is subtracted from every model and refined output. After this correction, the engagement gains become smaller but remain significant for all seven engagement metrics, the fluency trade-off remains as well, and eight of the nine dataset $\times$ model combinations keep the engagement-up/fluency-down pattern (Table 11 in the Appendix). The one exception is Llama-3.1 on ELI5, where FED scores track surface features most strongly and the correction reverses the engagement gain; overall, the bulk of the refinement signal is not explained by the surface features that drive FED’s spoken-vs-written gap.

6 Conclusion

We conducted a comprehensive study of the dialogue completion task in three explanation domains (human-spoken, human-written, and synthetic), regarding the task as a probe for both LLM and metric behavior. We found LLMs capable of efficiently formulating knowledge for explanatory dialogues, yet often inadequate to replace human explanations owing to poor interactivity. The outcome highly depends on the provided dialogue context: prompt phrasing has only a small effect, while trimming the context significantly lowers explanation quality.

According to two reference-free test suites, LLMs struggle with the interplay with the interlocutor, especially at dialogue level, and fall short of the baseline on human-spoken conversations, likely owing to the disparity between colloquial and written language. Nevertheless, LLMs are proven capable of refining their own output through metric-driven feedback, from which even smaller models could already benefit.

argument concerns the feedback signal, not the revising LLM.

Limitations

The **datasets** adopted in this work are rather small, especially ReWIRED. Considering the task accomplishment rate, the final amount of data points per batch sometimes falls below 400, 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 iterative self-refinement 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.

Our claims about engagement, interactivity, and refinement gains rest on **automatic evaluation without human judgments**. We triangulate between three independent signals – FED, IXQUISITE-only feedback, and LLM-as-a-judge preference from two model families – and they converge on the same conclusions, but convergence across automatic evaluators does not fully substitute for human validation of perceived dialogue quality. A targeted replication of the 300-turn preference protocol (§5.2) with a panel of human raters, reporting agreement against the two LLM judges, is left for future work.

Use of AI Assistants

Beyond the LLMs that are the object of study in our experiments, we used an AI coding assistant (Claude Code, Anthropic) to help implement the statistical analysis scripts accompanying the

camera-ready revision and to assist with L^AT_EX editing. All AI-assisted code and text were reviewed and verified by the authors, who take full responsibility for the content of this paper.

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.
- Aliki Anagnostopoulou, Nils Feldhus, Yi-Sheng Hsu, Milad Alshomary, Henning Wachsmuth, and Daniel Sonntag. 2025. [Human and LLM-based assessment of teaching acts in expert-led explanatory dialogues](#). In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 166–181, Suzhou, China. Association for Computational Linguistics.
- 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.
- Dan Carpenter, Wookhee Min, Seung Y. Lee, Gamze Ozogul, Xiaoying Zheng, and James C. Lester. 2024. [Assessing student explanations with large language models using fine-tuning and few-shot learning](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2024, Mexico City, Mexico, June 20, 2024*, pages 403–413. Association for Computational Linguistics.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. [PLACES: prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 814–838. Association for Computational Linguistics.
- 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.
- 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, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhui 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.
- 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 Yujia 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*.
- Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. [Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors](#). 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 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.
- Shikib Mehri and Maxine Esk  nazi. 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 Esk  nazi. 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 Pandealea, 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 1: 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*:

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.

Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

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

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.

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

TOPIC: Black hole **LEVEL:** Child

EXPERT: So have you ever heard of something called a black hole?

CHILD: What is a black hole?

EXPERT: Well, it has to do with, a lot with gravity, do you know what gravity is?

CHILD: No, not at all.

EXPERT: It's what keeps us on the earth.

CHILD: What?

EXPERT: 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.

CHILD: Nice.

EXPERT: The main thing about black hole to remember is that it's just, like I said, how the earth holds you down, the black hole pulls you in, as well. Now, try and take the ball from me, black hole--

CHILD: Why do you hold it so tight?

EXPERT: 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.

:

TOPIC: Black hole **LEVEL:** Colleague (another expert)

COLLEAGUE: So how do you do your observations in optical and infrared?

EXPERT: So fortunately there's, I'm also doing it from space with the Spitzer Space Telescope, so particularly in the infrared, and my main interest has been to try and study the environment around the super massive black holes, not as close as where the X-rays are coming from, but clearly there's something from the X-ray corona that illuminates the rest of the accretion disk, and the dust that's further out. [...] And so that, I love that, the ability to exchange time for resolution, because these structures are so far away that we're never gonna get a telescope big enough where that has the resolution to see the accretion disk, or the dust distribution around--

COLLEAGUE: So do you get dimensions of the disk out of that?

EXPERT: Yeah, again, we don't know exactly where X, Y, Z, zero is, we're assuming that it's something, you know, the X-rays that are coming out are very close to the event horizon of the black hole, but this is still, you know, your realm of X-rays, to really figure out those kinds of things. But once the X-rays, once the photons hit the corona, and are re-scattered, and up, energized, [...] this bright flash go off in a nearby AGN called NGC5548, and then you see it propagate as it warms up the disk, as all that light is falling onto it, and then eventually you hit the, the, further away, where the dust is, and the dust tends to radiate in an infrared. So we got basically a structure, and you just, you see this flashbulb go off, and then it illuminates, effectively, the structure.

CHILD: So you can map out the dust, where do you see it?

EXPERT: So you do see it, basically the dust sublimation radius, and you [...]

:

ANNOTATIONS

T01	D01	E02	Topic 01: Main topic
T01	D02	E04	Topic 03: A related topic
T03	D01	E01	
D08	E07		Dialogue Act 01: To ask a check question
D09	E03		Dialogue Act 02: To ask how/what question
D02	E06		Dialogue Act 03: To ask other kind of questions
T03	D09	E03	Dialogue Act 04: To answer a question by confirming
D07	E05		
T01	D09	E03	Dialogue Act 07: To provide agreement statement
D03	E04		Dialogue Act 08: To provide disagreement statement
T01	D09	E03	Dialogue Act 09: To provide informing statement

ANNOTATIONS

T03	D02	E02	Explanation Act 01: Testing understanding
			Explanation Act 02: Testing prior knowledge
T01	D09	E03	Explanation Act 03: Providing an explanation
			Explanation Act 04: Ask for an explanation
D01	E04		Explanation Act 05: Signalling understanding
			Explanation Act 06: Signalling not understanding
T03	D04	E03	Explanation Act 07: Providing feedback

D03 E04

T01 D09 E03

Figure 7: Two excerpts of explanatory dialogues from the WIRED dataset (Wachsmuth and Alshomary, 2022), exemplifying explaining the same topic (black hole) to different explainees (child and expert). 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.

Evaluated Aspect	ReWIRED	WikiDialog	ELI5
Relevant	6.92	7.40	7.78
Interesting	-11.13	-15.20	-9.03
Fluent	9.90	12.77	8.20
Correct	6.94	7.46	7.81
Understandable	9.06	11.95	7.66
Semantically appr.	10.44	13.10	8.41
Specific	-7.15	-11.02	-5.98
Engaging	-14.14	-18.22	-11.45
Likeable	-8.33	-12.36	-6.65
Understanding	6.91	7.41	7.78
Coherent	6.90	7.40	7.76
Topic depth	6.95	7.47	7.80
Flexible	-3.37	-4.62	-3.00
Diverse	6.99	7.51	7.84
Inquisitive	-12.44	-18.64	-9.57
Error recovery	6.84	7.31	7.72
Informative	-6.73	-5.88	-4.79
Consistent	7.02	7.56	7.86
Minimal explanations	0.16	0.85	0.20
Lexical complexity	1.15	2.42	1.52
Synonym density	0.09	0.00*	0.00*
Adaptation	0.24	0.15	0.20
Readability level	0.46	0.65	0.45
Coherence	0.04	0.03	0.03

Table 6: The differences between dialogues from the three datasets according to the two test suites. The best scores for each aspect is highlighted in boldface. The starred zeros result from limitations of the datasets.

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 7). Moreover, the dataset features turn-level labels that highlight interactions categorized into dialogue acts, and explanation acts. The expanded ReWIRED corpus in [Anagnostopoulou et al. \(2025\)](#) 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.

While using the full ReWIRED dataset, we **re-sample** WikiDialog and ELI5 through random selection of respectively 2,000 and 1,000 dialogues. Table 13 to 15 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.

A.2 Dataset quality (baseline)

We measure the baseline for each dataset by how FED (Table 5) and IXQUISITE (Table 7) rate the original turn. The scores are shown in Table 6. 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.

Among the IXQUISITE aspects (Table 6), “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.

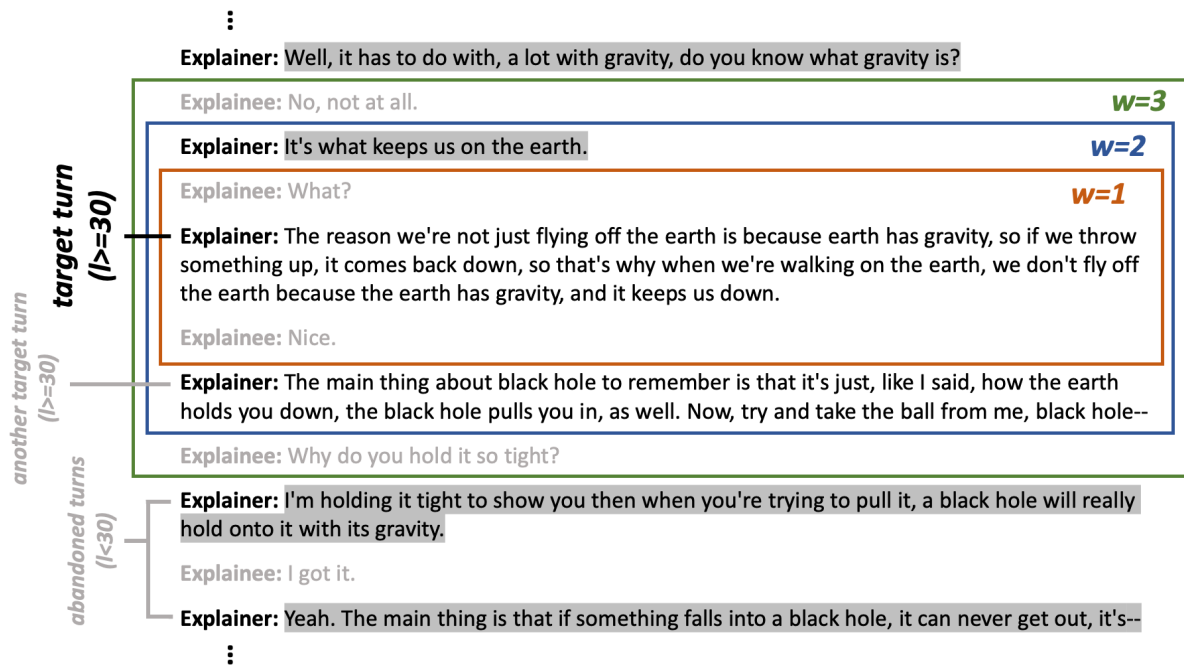


Figure 8: Example from the WIRED dataset of how data instances are built from a turn and its surrounding dialogue context. Minimum turn length l is set to 30 here.

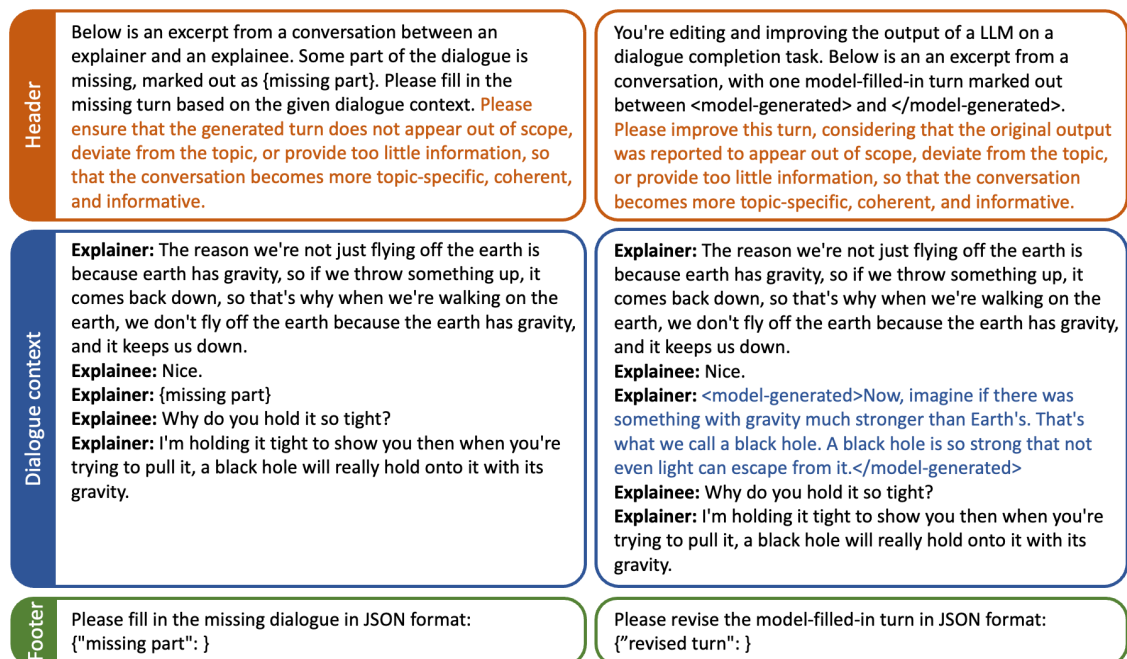


Figure 9: Two prompt structures for iterative self-refinement: the original prompt with instruction (left) and the rewritten header (right). The instructions are highlighted in both header columns.

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 7: The 6 numerical measures in IXQUISITE (Anagnostopoulou et al., 2025) adopted in this work. Correlation to teaching model from the original test suite is left out because of the lack of annotations.

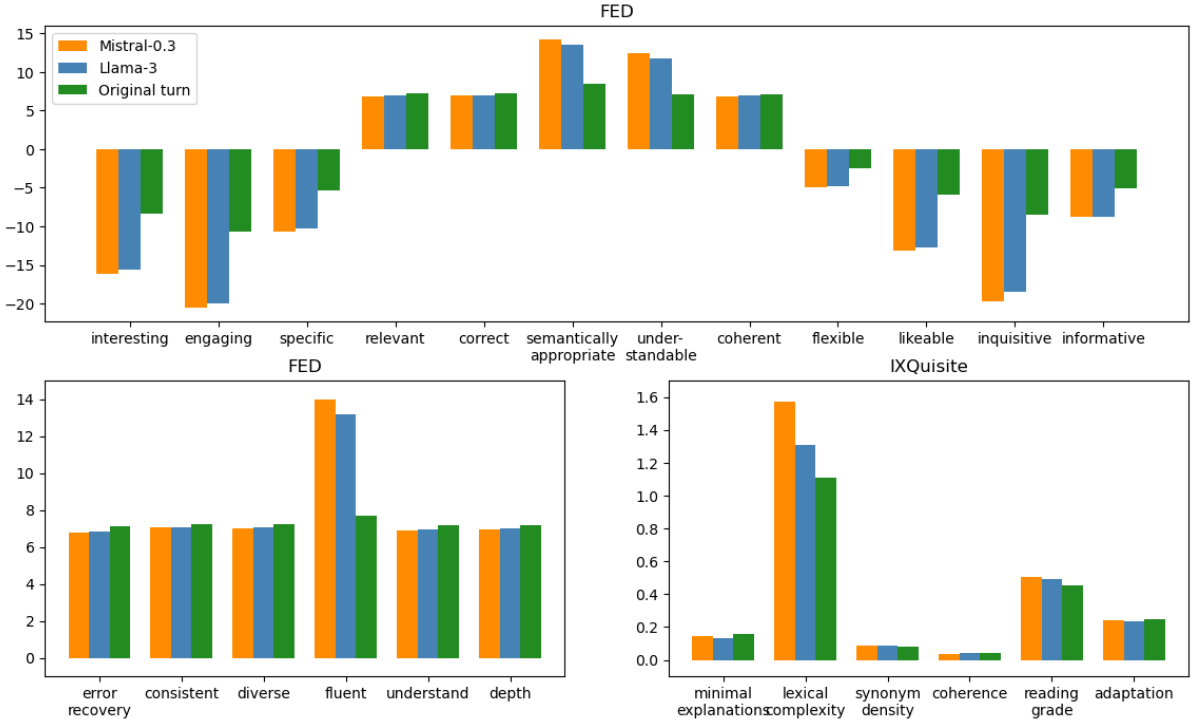


Figure 10: Evaluation results of the LLM-generated and the original human turns measured by FED and IXQUISITE in the trial run.

B Prompt variables

We define four variables to form the prompt differently: *topic*, *speakers*, *open-end*, and *footer w/ context*. While *topic* and *speakers* offer additional information to the given dialogue segment, *open-end* and *footer w/ context* structurally alters how the dialogue context is established or provided to the model.

Additionally, Figure 8 visualizes how window w is applied to control the prompted dialogue context. Figure 9 compares the two prompting strategies in iterative self-refinement: adapting the original task prompt vs. establishing a new prompt header for refinement.

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.

Footer w/ context. The variable defines whether the dialogue context is prompted again while demanding an output in JSON format. Such design

purposes the models to generate continuations of the dialogue instead of filling in the gap between turns.

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

C 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 10), 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 11) and Llama-3 (Figure 12) 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.

Table 16 demonstrates a comparison between task output with and without *open-end* prompt variable, which is found to impact model output the most. Furthermore, this effect extends to the subsequent stages of the experiment. Table 17 sheds light to how iterative self-refinement 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.

D Supplementary results

Task performance. Extending the contents in Figure 4, Figure 13 and Figure 14 show the task performance of WikiDialog and ELI5. In the visualization, the four bars of each model represents a prompt variant: *vanilla*, *topic & speakers*, *open-end*, and *topic & speakers & open-end*. The black box denotes baseline, i.e., the original turns.

Worst-performing features. As an extension of Table 2, Figure 15 and Figure 16 demonstrate the counts for worst-performing features that are retrieved for iterative self-refinement.

Iterative self-refinement. We tested two prompting strategies for specifying the identified worst features. As shown in Figure 17, we discover that establishing a refinement-specific prompt results in more significant improvement than adapting the original prompt. This trend is universal across the other prompt variables.

Echoing the framework by Madaan et al. (2023), iterative self-refinement consistently leads to improvement of the dialogue completion task performance. Across two prompt variants (*vanilla* and *open-end*), the results are respectively visualized for each dataset: ReWIRED (Figure 18), WikiDialog (Figure 19), and ELI5-dialogues (Figure 20). 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 self-refined (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.

E Bias diagnostics and ablation details

This appendix collects the detailed statistics behind the bias diagnostics and robustness checks reported in §5.2 and §5.4. Unless noted otherwise, significance refers to paired t-tests at $p < 0.05$.

FED metric	len	word len	punct.	TTR
Engaging	+0.58	-0.27	+0.09	-0.42
Interesting	+0.60	-0.27	+0.16	-0.42
Likeable	+0.58	-0.28	+0.08	-0.42
Informative	+0.46	-0.07	+0.17	-0.29
Fluent	-0.57	+0.27	-0.22	+0.38
Understandable	-0.54	+0.24	-0.23	+0.36

Table 8: Pearson r between surface features of the human reference turn (token count, mean word length, punctuation density, type–token ratio) and its FED score, pooled over $N=4,273$ reference turns from all three datasets. All correlations significant at $p < 0.01$; boldface marks $|r| \geq 0.20$.

Dataset	Judge	% human	% task	% tuned
ReWIRED	Gemini 2.5 Pro	65.0	22.0	13.0
	Claude Haiku 4.5	85.0	7.0	8.0
WikiDialog	Gemini 2.5 Pro	46.0	23.0	31.0
	Claude Haiku 4.5	58.0	19.0	23.0
ELI5	Gemini 2.5 Pro	66.0	16.0	18.0
	Claude Haiku 4.5	62.6	14.1	23.2

Table 9: LLM-as-a-judge preference rates per dataset ($N=100$ turns each) for the two judges on the blind 3-way ranking over {human, task, tuned} turns.

Surface-feature regression. Table 8 reports the Pearson correlations between the four surface features and the FED scores of the human reference turns, pooled over $N=4,273$ turns (521 ReWIRED, 2,044 WikiDialog, 1,708 ELI5).

LLM-as-a-judge details. The 300-turn sample is drawn length-stratified with a fixed seed, and the A/B/C answer labels are shuffled per row with the same seed for both judges (gemini-2.5-pro and claude-haiku-4-5-20251001). Table 9 reports per-dataset preference rates. On the $N=299$ rows completed by both judges, 3-way exact agreement is 60.9% (Cohen’s $\kappa=0.262$), and agreement on the binary “human-best” verdict is $\kappa=0.311$ – fair agreement for a 3-way categorical task with class imbalance.

IXQuisite-only ablation. Table 10 compares the paper’s refinement setup (FED+IXQUISITE feedback) with the ablation that uses only IXQUISITE features as feedback, both evaluated against the un-refined task output on Claude-3 $\times$ ELI5 with the vanilla prompt ($N=1,225$). Under IXQUISITE-only feedback, all seven engagement metrics improve significantly at $p < 0.001$; under FED+IXQUISITE feedback, six of seven do (“flexible” falls short of significance).

Bias-corrected refinement effects. For each FED metric, a linear regression fitted on the hu-

Cluster	FED+IXQ feedback		IXQ-only feedback	
	sig.	mean d	sig.	mean d
Engagement (7)	6/7 $\uparrow$	+0.41	7/7 $\uparrow$	+0.22
Fluency (3)	3/3 $\downarrow$	-0.41	3/3 $\downarrow$	-0.24
Other (8)	0/8	± 0.00	8/8 $\uparrow$	+0.77

Table 10: IXQuisite-only refinement ablation on Claude-3 $\times$ ELI5 (vanilla prompt, $N=1,225$): number of significantly changed metrics and mean Cohen’s d per cluster, relative to the un-refined task output.

Dataset	Model	N	Engagement d		Fluency d	
			raw	corr.	raw	corr.
ReWIRED	Llama-3.1	482	+0.51	+0.31	-0.49	-0.27
	Mistral-0.3	192	+0.40	+0.23	-0.42	-0.15
	Claude-3	242	+0.67	+0.51	-0.68	-0.51
WikiDialog	Llama-3.1	1,861	+0.64	+0.39	-0.80	-0.55
	Mistral-0.3	800	+0.41	+0.32	-0.51	-0.40
	Claude-3	1,170	+0.36	+0.17	-0.52	-0.36
ELI5	Llama-3.1	1,583	+0.42	-0.14	-0.45	+0.18
	Mistral-0.3	632	+0.17	+0.12	-0.16	-0.08
	Claude-3	199	+0.41	+0.29	-0.41	-0.29

Table 11: Mean Cohen’s d (refined vs. task output) for the engagement and fluency clusters per dataset $\times$ model cell, on the raw FED scores and after subtracting the surface-feature contribution (corr.).

man reference turns only ($N=4,273$) predicts the score from turn length, mean word length, punctuation density, and type–token ratio; this prediction is subtracted from the score of every model output and refined output. Table 11 reports the per-cell engagement and fluency effect sizes before and after the correction. Pooled across all nine cells ($N=7,161$ paired turns), the mean engagement effect goes from $d=+0.452$ to $+0.214$ and the mean fluency effect from -0.538 to -0.286 , with all seven engagement metrics remaining individually significant.

F Reproducibility details

Data splits. Our study is evaluation-only: no model is trained or fine-tuned, so no train/development/test split applies. Dataset sizes, filtering steps, and sampling are reported in Appendix A.1, and per-dataset instance counts in Table 1.

Inference settings. The open LLMs run locally with transformers, using sampling with each model’s default generation configuration and a limit of 512 new tokens. Claude-3 Haiku is queried through the Anthropic API with a limit of 100 output tokens, which explains the truncation-related task failures discussed in §5.1. The LLM-as-a-judge study uses gemini-2.5-pro and claude-haiku-4-5-20251001 with a fixed sampling seed and JSON response format. We did

not perform hyperparameter search: the instance-building variables were chosen based on the trial run (Appendix C, $l=30$, $w=2$), and the number of worst-performing features ($n=3$) was set to keep instructions short (§3.4).

Reporting. All scores come from a single inference run per configuration. Reported values are means over all data points of a configuration, and statistical claims are supported by paired t-tests with Cohen’s d as effect size.

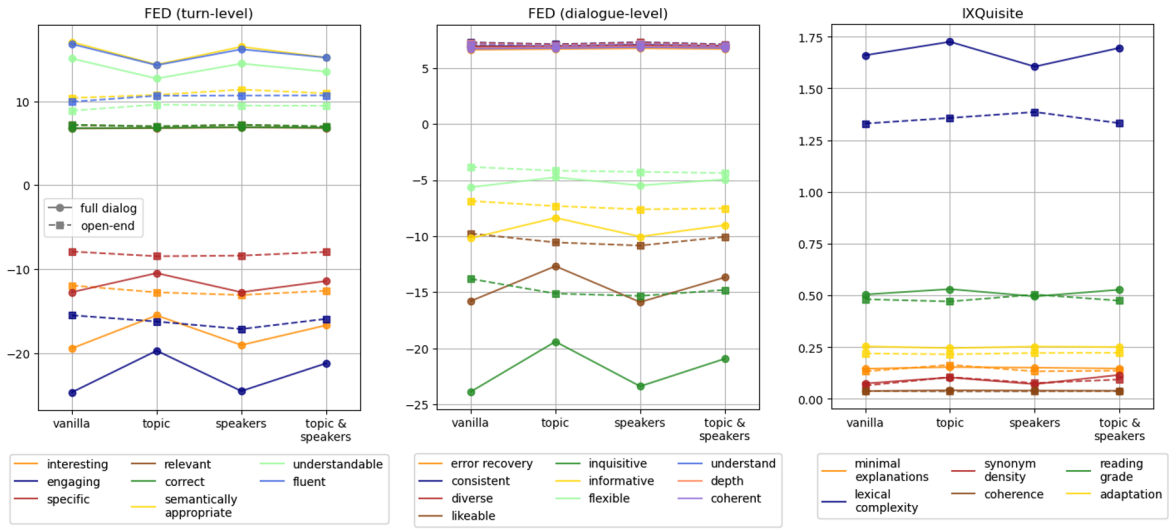


Figure 11: The different task performance of Mistral-0.3 across three variables in prompt design: *topic*, *speakers*, and *open-end*.

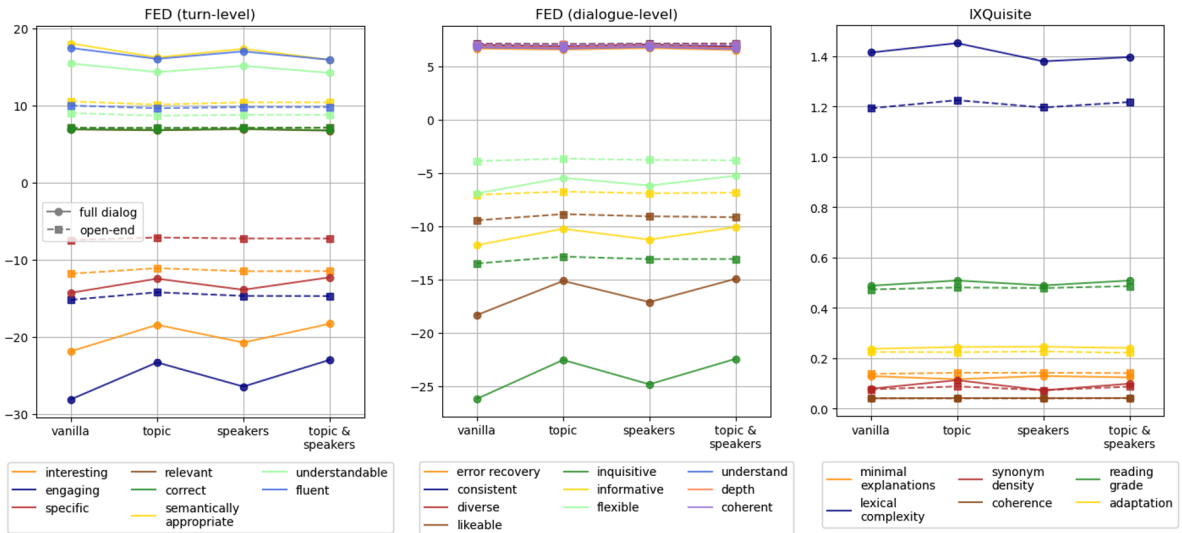


Figure 12: The different task performance of Llama-3 across three variables in prompt design: *topic*, *speakers*, and *open-end*.

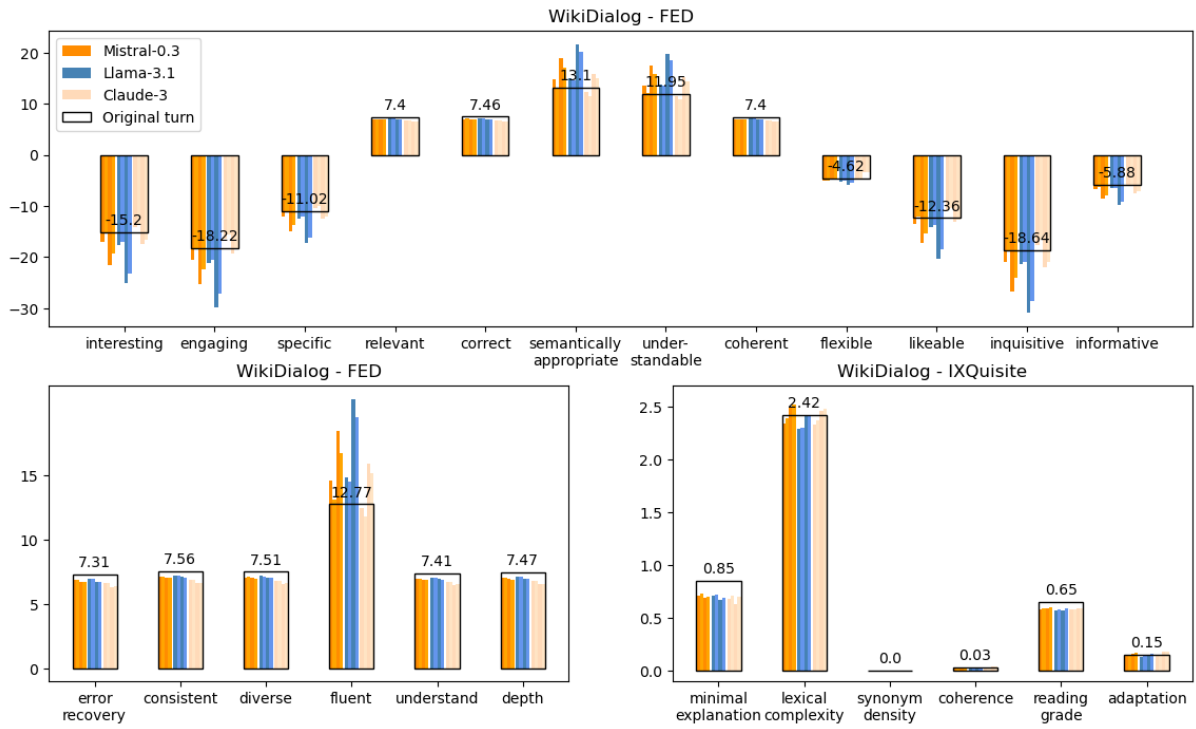


Figure 13: Evaluation of the LLM-generated and the original human turns with FED (top and bottom-left) and IXQUISITE (bottom-right) on WikiDialog.

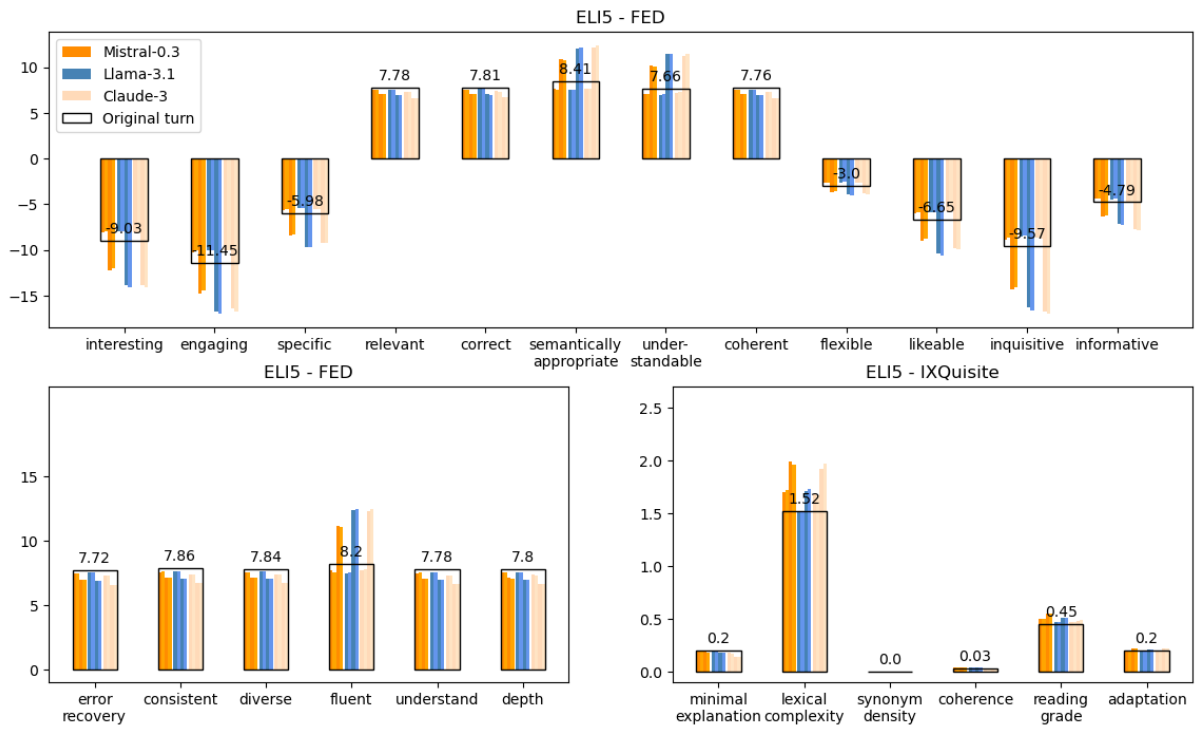


Figure 14: Evaluation of the LLM-generated and the original human turns with FED (top and bottom-left) and IXQUISITE (bottom-right) on ELI5.

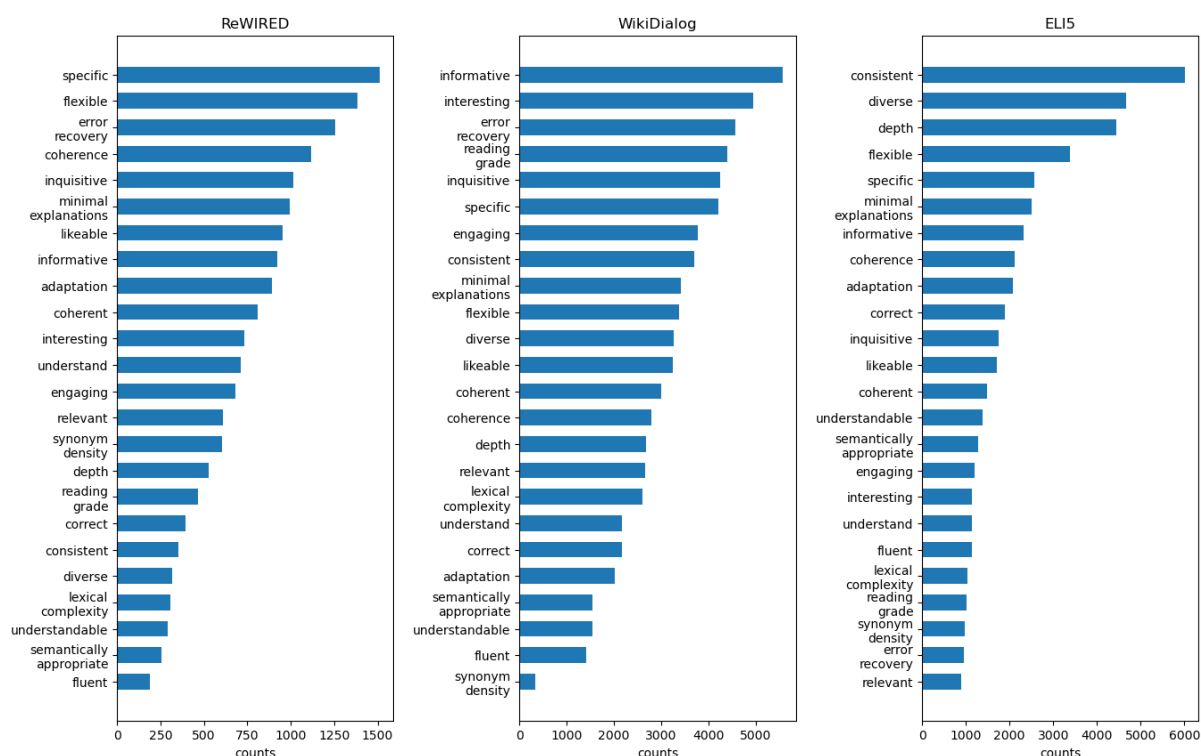


Figure 15: The worst performing features after normalization for each dataset.

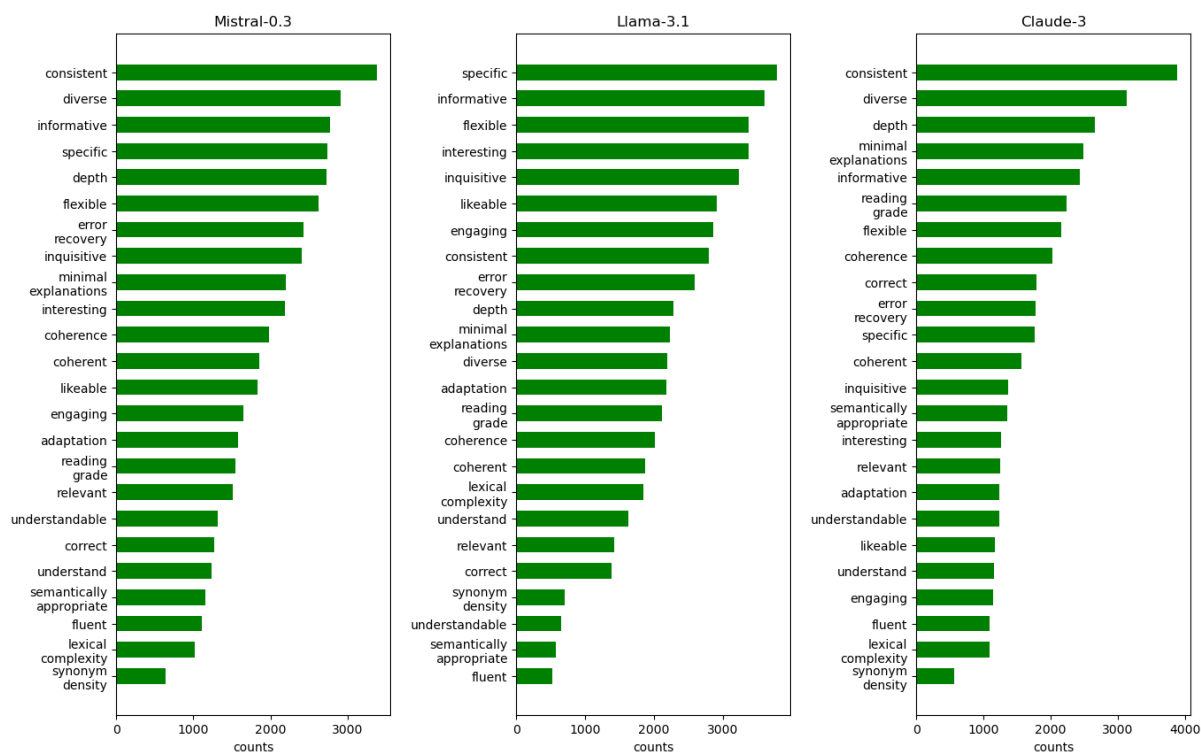


Figure 16: The worst performing features after normalization for each LLM.

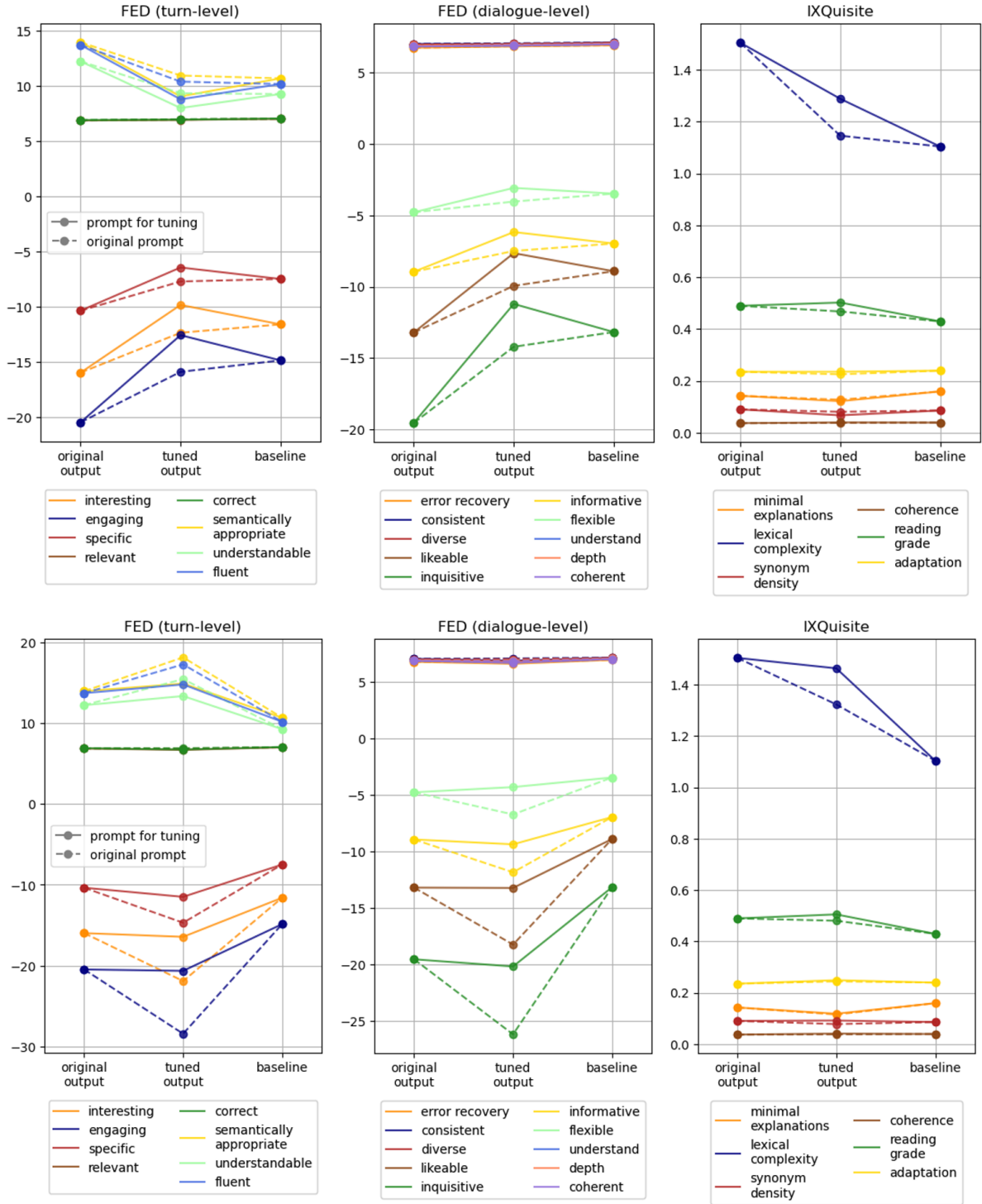


Figure 17: The performance improvement of and Llama-3.1 under iterative self-refinement with two different prompting strategies, i.e., designated prompt for refinement vs. adapting the original task prompt. Task output of two prompt variants are tested: vanilla (top) and the *topic & speakers & open-end* prompt (bottom).

ReWIRED

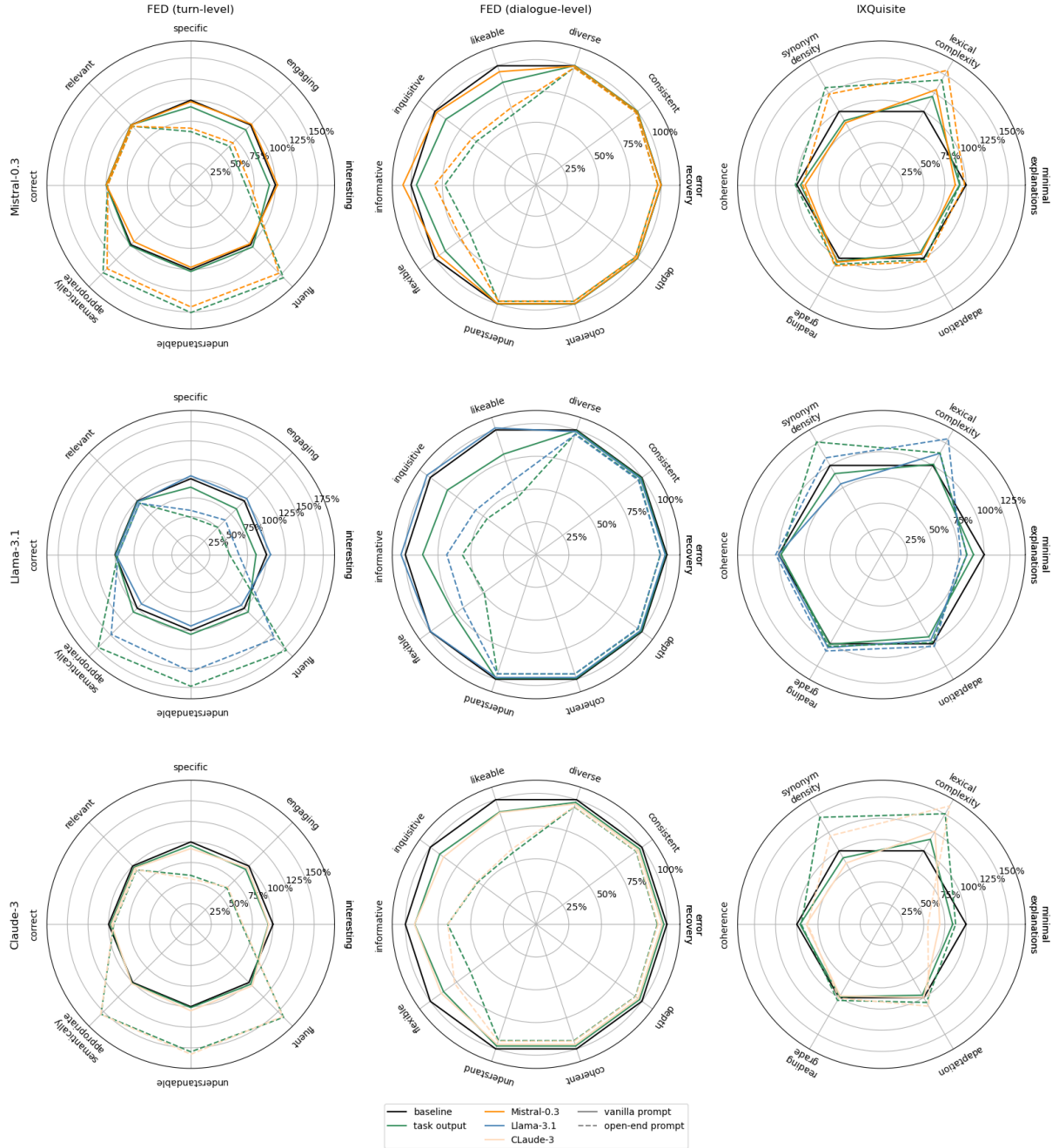


Figure 18: The evaluation results of refined turns of the ReWIRED dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the colored lines indicate the refined output for each model. Solid and dashed lines respectively represent vanilla and *open-end* prompt.

WikiDialog

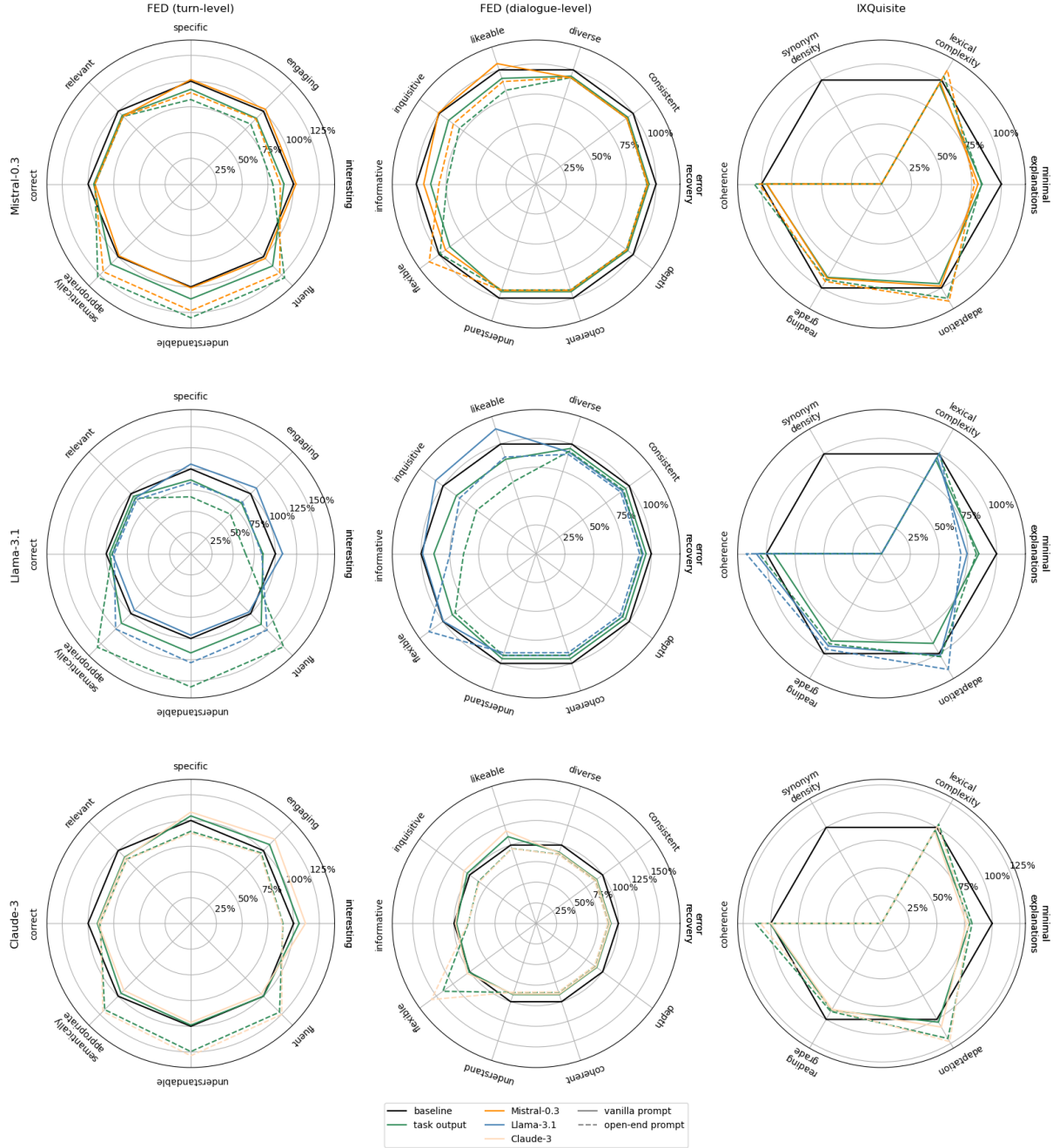


Figure 19: The evaluation results of refined turns of the WikiDialog dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the coloured lines indicate the refined output for each model. Solid and dashed lines respectively represent vanilla and *open-end* prompt. Since dialogue topics from the dataset are often proper nouns, “synonym density” in IXQUISITE is left out in plotting.

ELI5-dialogues

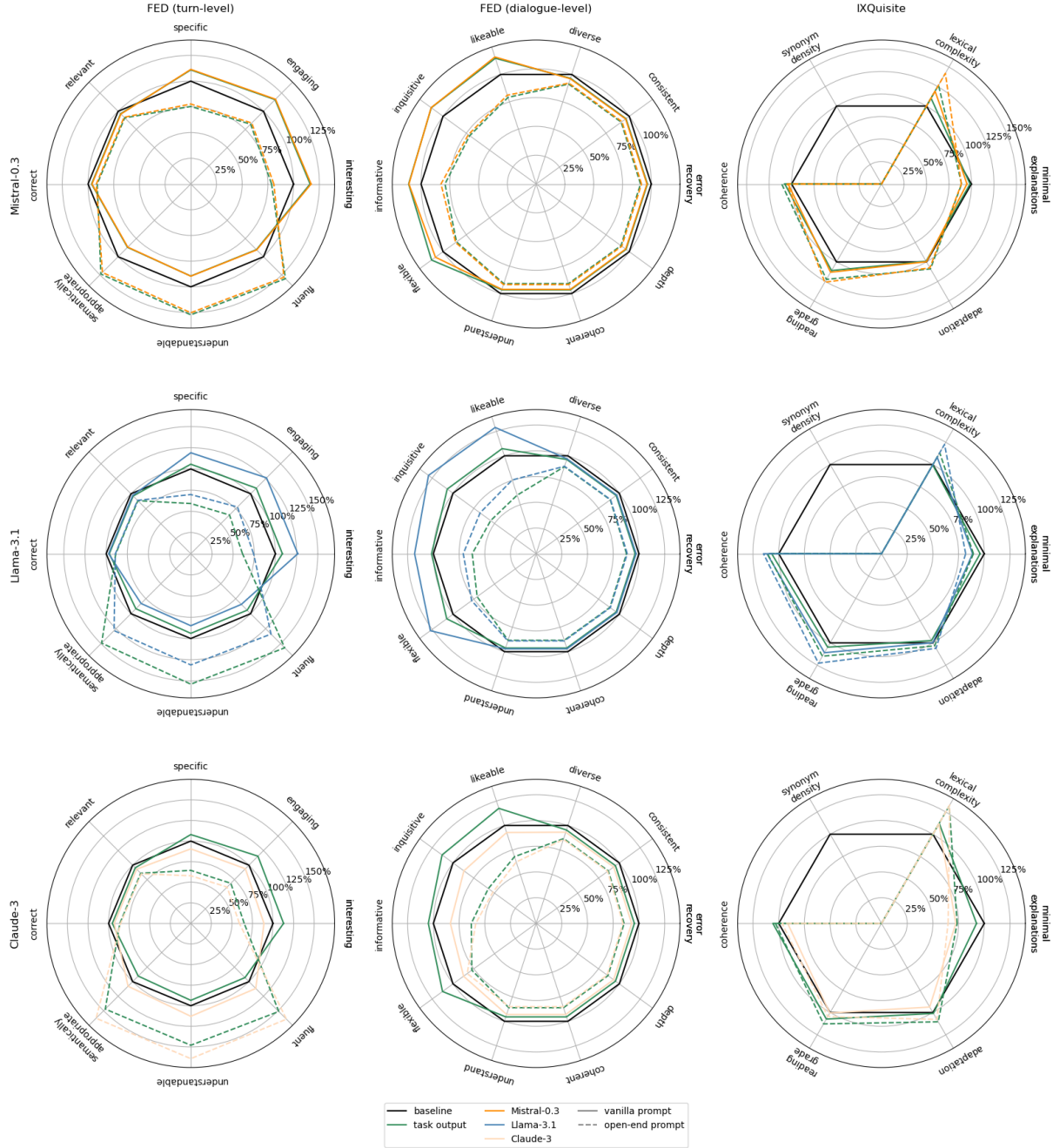


Figure 20: The evaluation results of refined turns of the ELI5 dataset. The black lines (100%) represent the baseline of the original turns. The green lines denote the original output, and the colored lines indicate the refined output for each model. Solid and dashed lines respectively represent vanilla and *open-end* prompt. Owing to the lack of topic from the dataset, “synonym density” in IXQUISITE is left out in plotting.

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 12: The evaluated aspects and their corresponding negative and positive description for iterative self-refinement. 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	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 13: 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 kënnen mir de Toun an dem Himmel, wenn du mie 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	Explainer: What was the best mark of the song? Explainee: Modern Times also recorded the song in English, as "Heaven Only Knows."		

Table 14: 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 15: 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, <i>gravity is a result of distortion of spacetime</i> by the presence of mass. So <i>when the train in your example accelerates</i> , 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 <i>the geometry of spacetime</i> . 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 <i>time dilation and length contraction</i> .
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 16: 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	Refined 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 17: Example of a comparison between task output and self-refined 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.