Meta-Prompting Follow-Ups for Unsupervised Dialogue Evaluation

Anonymous ACL submission

Abstract

Automatically evaluating dialogue quality poses considerable challenges due to the complex nature of conversational interactions. In this paper, we introduce DIET, a new unsupervised and reference-free metric designed to enhance dialogue evaluation. DIET uses domain-specific follow-up utterances generated through meta-prompting with Large Language Models (LLMs). We propose two meta-prompting strategies: Self-Referential Follow-Up Meta-Prompting (SELFMAP), where the same model is responsible for both generating and evaluating the follow-ups, and Cross-Model Follow-Up Meta-Prompting (CRAFT), which uses an ensemble of different models. Dialogue quality is then measured by determining the likelihood that an LLM would continue the conversation based on the generated follow-ups. We assess DIET's correlation with human judgments and its performance in discourse parsing, extracting dialogue structures. Experiments show DIET outperforms existing reference-free metrics across datasets.

1 Introduction

Human evaluation is often considered the best way to assess dialogue quality due to its ability to grasp nuances and context. However, the high cost, lack of reproducibility, and inconsistency in human ratings make automatic metrics a necessary complement (Huang et al., 2020; Mehri et al., 2022). Despite advancements in Natural Language Processing (NLP), creating a reliable automatic metric for dialogue evaluation remains difficult due to the involvement of multiple interlocutors and the complex dynamics of context (Zhao et al., 2017; Zhang et al., 2021b). Traditional metrics like BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), commonly used in language evaluation, struggle with dialogue tasks due to the variability in one-to-many conversational responses, leading to poor alignment with human judgments

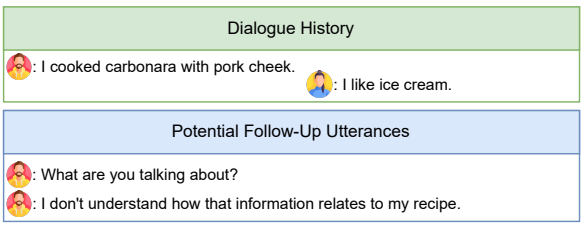

Figure 1: A dialogue history with potential follow-up utterances designed for evaluation purposes.

(Liu et al., 2016; Yeh et al., 2021). This problem is exacerbated in settings where reference responses are unavailable, such as online evaluations. Furthermore, the scarcity of high-quality annotated dialogue data poses a significant challenge to the development of effective supervised metrics. Therefore, there is a significant need for unsupervised metrics that can evaluate dialogues accurately without relying on extensive annotated datasets.

Evaluating the quality of a dialogue can be effectively conducted through the analysis of follow-up utterances. This approach, as demonstrated by Eskénazi et al. (2019), leverages the implicit feedback embedded within user responses to assess the interaction. When users engage in a conversation, their subsequent utterances often reflect their perceptions of the conversation's quality. For instance, consider the dialogue depicted in Figure 1. In this scenario, a concluding remark such as "*What are you talking about?*" indicates confusion or dissatisfaction, signaling that the user's understanding or engagement with the conversation has faltered. This type of response suggests a breakdown in communication, where the user is struggling to follow or find relevance in the interaction. Alternatively, a more pertinent follow-up within the food domain might be "*I don't understand how that information relates to my recipe*". This response shows a more specific issue, indicating that while the user is engaged, they are seeking clearer connections

between the provided information and their immediate context or needs. Recent approaches, such as FED (Mehri and Eskénazi, 2020) and FULL (De Bruyn et al., 2022), have built on this idea by using follow-ups to assess dialogue quality. Specifically, these methods involve querying a language model to simulate human responses and then measuring the likelihood of generating a fixed set of follow-ups. While promising, these metrics have their own limitations. For instance, both metrics define follow-ups using annotated data. FED applies minimal tuning on a small sample of conversations, while FULL determines follow-ups through correlation analysis with annotated data. These supervised methods may bias results toward specific datasets, limiting their applicability across domains (Kawamoto et al., 2023).

In this paper, we introduce **DIET**[1] (**DI**alogue **E**valuation via me**T**a-prompting), a novel unsupervised metric designed to enhance dialogue evaluation without relying on reference responses. DIET leverages domain-specific follow-ups generated through meta-prompting techniques with Large Language Models (LLMs) (de Wynter et al., 2023). Meta-prompting aims to automatically create prompts that help LLMs generate contextually relevant and insightful follow-ups tailored to the dialogue. This method is inspired by approaches that reduce the need for extensive prompt engineering in NLP (Hou et al., 2022; Suzgun and Kalai, 2024) and incorporates concepts of metacognition, where the model essentially "*thinks about its thinking*", to foster self-awareness and self-regulation in LLMs (Wang and Zhao, 2023). Two meta-prompting strategies are proposed: Self-Referential Follow-Up Meta-Prompting (SELFMAP), where the same model generates and evaluates the follow-ups, ensuring that the responses are within the model's capabilities; and Cross-Model Follow-Up Meta-Prompting (CRAFT), which uses an ensemble of different models to generate a set of follow-ups, leveraging the strengths and mitigating the biases of individual models. By automating follow-up selection, DIET overcomes the limitations of existing metrics like FED and FULL by broadening its application from open-domain conversations to goal-oriented dialogues.

We evaluate DIET on two streams of tasks: intrinsically on dialogue quality evaluation and extrinsically on dialogue discourse parsing. On dialogue quality, DIET outperforms FED and FULL across two datasets, indicating that joint log-likelihood offers a more effective alignment for assessing open-domain dialogues, irrespective of the meta-prompting strategy adopted. In dialogue discourse parsing, DIET exceeds FED and FULL in the evaluation of dialogues to extract their structural elements. This improvement is due to the use of conditional log-likelihood and the CRAFT strategy, which yields a more comprehensive and precise set of domain-specific follow-ups.

The main contributions of this paper include:

- DIET, a new unsupervised, reference-free metric for dialogue quality evaluation.
- Two LLM-based meta-prompting strategies to auto-generate domain-specific follow-ups with minimal human intervention.
- DIET's improved alignment with human judgments over FED and FULL across benchmark datasets.
- Superior performance in discourse parsing, showing the effectiveness of domain-specific follow-ups for accurate dialogue assessment.

2 Related Work

Dialogue evaluation metrics fall into two main categories: reference-based and reference-free. The latter includes follow-up methods like the one proposed in this paper.

Reference-Based Metrics Reference-based metrics evaluate generated responses against reference utterances using traditional word-overlap metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), Diversity (Li et al., 2016), and Entropy (Zhang et al., 2018). However, these metrics often misalign with human judgment due to the one-to-many nature of dialogues (Liu et al., 2016; Yeh et al., 2021). This has led to the adoption of learning-based metrics like ADEM (Lowe et al., 2017), RUBER (Tao et al., 2018), BERT-RUBER (Ghazarian et al., 2019), BERTScore (Zhang et al., 2020b), BLEURT (Sellam et al., 2020), and Deep AM-FM (Zhang et al., 2020a), which focus on semantic similarity and context over word overlap, aligning better with human evaluations. To address the shortcomings of single-reference comparisons,

[1]The name highlights a key advantage of our metric compared to FED and FULL: it does not require annotated data to define follow-ups.

multiple reference responses have been proposed (Galley et al., 2015; Sordoni et al., 2015; Gupta et al., 2019). However, reference-based metrics struggle in practical use, as they are unsuitable for online evaluation, where reference responses are unavailable, and require resource-intensive human annotation for high-quality responses.

Reference-Free Metrics Reference-free metrics aim to assess dialogue quality without relying on human-defined references. For instance, Zhang et al. (2021a) introduced DynaEval, a framework using graph convolutional networks to assess dialogues modeled as graphs. Their training process involves generating negative samples by altering utterances or speaker order. Similarly, Park et al. (2021) introduced a method for generating negative examples by altering golden responses to produce inappropriate counterparts. This approach is employed to train an unsupervised response evaluation model, enhancing the model's ability to differentiate between appropriate and inappropriate responses. In a related approach, Mendonça et al. (2022) proposed QualityAdapt, which uses the Adapter paradigm (Houlsby et al., 2019) to train individual adapters on dialogue subqualities, with negative samples created through word reordering, dropping, and repetition. Conversely, Park et al. (2023) developed DENSITY, which measures the density of a generated response against the feature distribution of human responses by training a response selection model to distinguish correct responses from randomly sampled negative examples. However, these methods struggle with artificial negative samples, failing to capture the nuanced incoherence in human conversations (Ghazarian et al., 2022). Alternative approaches use Abstract Meaning Representation (AMR) graphs to create natural incoherence for training evaluation models (Ghazarian et al., 2022) or model core semantic information (Li et al., 2024b), but AMR parsers may introduce errors in graph generation.

All the mentioned metrics are supervised and rely on costly, expert-annotated data. To address this, research has shifted to unsupervised metrics using Pre-trained Language Models or LLMs. For instance, Dial-M evaluates dialogues by masking keywords and using cross-entropy loss (Dey and Desarkar, 2023), though it may still favor incoherent responses if common phrases are masked. Recent studies have also investigated the performance of LLMs in dialogue evaluation tasks (Lin and Chen, 2023; Chan et al., 2024; Zhang et al., 2024). Nevertheless, findings indicate that LLMs may struggle to concentrate on relevant segments of dialogues and that their effectiveness is highly influenced by the prompt design (Finch et al., 2023; Jia et al., 2024).

Reference-Free Metrics based on Follow-Ups FED (Mehri and Eskénazi, 2020) and FULL (De Bruyn et al., 2022) are metrics that assess dialogue quality by evaluating how likely a language model is to continue a conversation using a predefined set of follow-up utterances.

FED uses the DialoGPT model (Zhang et al., 2020c) to assess 18 fine-grained dialogue qualities, with positive and negative follow-ups manually defined for each quality. For instance, for the "*interesting*" quality, positive follow-ups might include "*Wow! Very interesting*", while negative ones could be "*That's really boring*". These follow-ups were tuned on 10 conversations from the FED dataset (Mehri and Eskénazi, 2020). Dialogue quality scores are derived from the joint log-likelihood of generating these follow-ups, conditioned on the dialogue history and the follow-up itself. An overall impression score is the average of turn-level or dialogue-level quality scores.

FULL builds on FED with key differences. Instead of joint log-likelihood for the whole conversation, FULL calculates the conditional log-likelihood of follow-ups using the Blender model (Roller et al., 2021), conditioned on the dialogue history, avoiding biases from the training dataset (Reddit, for DialoGPT) and not favoring longer conversations. Unlike FED, which uses manually defined follow-ups, FULL selects them through correlation analysis of the FED dataset and provides a single overall quality score rather than 18 fine-grained attributes. Results showed that follow-ups intended for specific attributes (e.g., engagingness) often correlated with unrelated attributes (e.g., correctness), leading FULL to focus on overall quality.

FED and FULL show promise for unsupervised dialogue evaluation but have limitations. While both metrics assess dialogue quality in an unsupervised manner, they rely on a supervised approach for defining follow-ups and use generic follow-ups that may not fit specific domains. In contrast, DIET automates the selection of domain-specific follow-ups without needing annotated data.

Another challenge is compositionality, defined in Mehri and Eskénazi (2020) as the ability of a

generative model to produce utterances that have never been observed. For instance, it should generate "*That doesn't make sense*" even if not explicitly trained with such a follow-up. While models like DialoGPT can generate some novel follow-ups, they struggle with less common or complex concepts like consistency and comprehensibility (Mehri and Eskénazi, 2020). DIET addresses this by using follow-ups familiar to the evaluation model, reducing reliance on compositionality.

3 Method

Our goal is to automate dialogue quality evaluation through follow-ups with minimal or no human intervention. To accomplish this, we reframe the task of defining follow-ups as a text generation problem, instructing LLMs to generate utterances using meta-prompting strategies. Here, meta-prompting creates prompts that guide language models to produce insightful, tailored follow-ups based on dialogue context and domain. Figure 2 depicts the DIET's methodology in three steps: (i) defining the meta-prompt, which includes domain-specific and general descriptions to guide LLMs in generating follow-ups; (ii) generating follow-ups using meta-prompting; and (iii) computing dialogue quality scores by analyzing the log-likelihoods of the dialogue history and generated follow-ups.

3.1 Meta-Prompt Definition

As depicted in Figure 2, the meta-prompt comprises three components: (i) a description of using follow-ups for dialogue evaluation D_{fue}; (ii) an in-context example I to guide the LLM towards the expected output type; and (iii) a domain-specific text D_τ, tailored to the domain τ and written by humans to provide domain knowledge for LLMs, along with formal instructions for generating follow-ups.

It is noteworthy that, although human involvement is still required in follow-up generation, it is significantly reduced. Specifically, D_τ is the only component of the meta-prompt that is domain-oriented and requires customization for each domain. In contrast, D_{fue} and I are generic and remain consistent across domains. The meta-prompt is then provided to LLMs to generates follow-ups. Ultimately, the dialogue history and generated utterances serve as inputs to assess a dialogue for domain τ. The meta-prompts designed for follow-up generation are presented in Appendix A.

3.2 Follow-Up Generation Strategies

To evaluate potential LLM biases, we propose two distinct meta-prompting strategies that differ in their choice of models for generating follow-ups and assessing dialogue quality.

SELFMAP This strategy addresses the compositionality issue by focusing on follow-ups a language model can generate. It seeks to ensure that the analysis of follow-ups relies on utterances within the model's generative capabilities, thereby minimizing dependence on the model's performance with novel or rare utterances. To achieve this, when choosing a language model for follow-up generation, referred to as LLM_1 in Figure 2, the strategy advocates for utilizing the same model during the dialogue quality evaluation step to assess utterances within the specified dialogue history.

To effectively implement this strategy, the LLM employed must be fine-tuned using instruction tuning (Zhang et al., 2023), which enhances its ability to follow specific directives and generate responses that align with given instructions, making it suitable for producing contextually appropriate and self-relevant follow-ups.

CRAFT Unlike SELFMAP, this strategy involves using different models for follow-up generation and dialogue quality evaluation. It is proposed for two reasons: (i) to assess whether the concept of compositionality (Mehri and Eskénazi, 2020) is necessary within the context of dialogue quality evaluation using LLMs, and (ii) to investigate if using multiple LLMs for generating follow-ups offers advantages over a single model. Each LLM has unique training and architecture, which enables each model to contribute distinct and potentially complementary perspectives in generating utterances. By leveraging multiple LLMs, we achieve a broader range of utterances, addressing aspects a single model might miss. Moreover, biases inherent in one model can be balanced by the utterances from other models, enhancing the diversity and richness of the follow-ups. Consequently, this strategy uses an ensemble of LLMs $(LLM_1, \ldots, LLM_m)$ for generating follow-up utterances, with each LLM receiving the same meta-prompt. The final set of follow-ups is formed by combining the outputs from these models, then a separate LLM (LLM_j, where $j \neq 1, \ldots, m$) is used for dialogue quality evaluation.

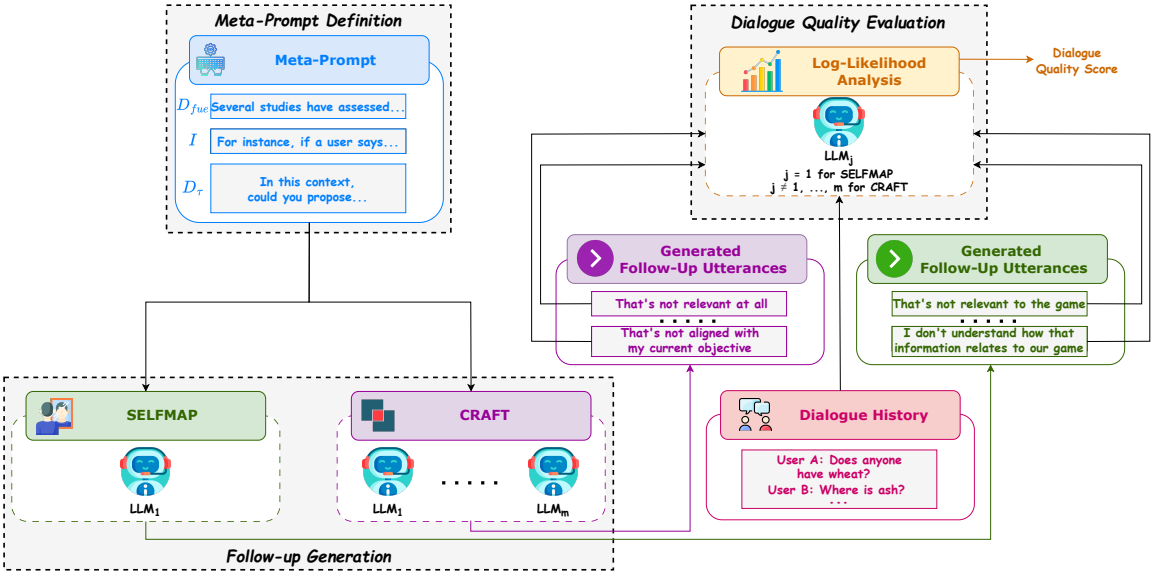

Figure 2: Overview of the DIET methodology, which involves: (i) Meta-prompt definition; (ii) Follow-up generation via SELFMAP and CRAFT; and (iii) Dialogue quality evaluation.

3.3 Dialogue Quality Evaluation via Log-Likelihood

DIET assesses the quality of a dialogue within a specific domain τ using an LLM and a set of n generated follow-ups $F_\tau = \{f_{1_\tau}, \ldots, f_{n_\tau}\}$. Following the FULL metric, we focus on the dialogue's overall quality, in contrast to fine-grained quality attributes proposed in FED. As depicted in Figure 2, for each follow-up utterance f_{i_τ}, DIET computes the log-likelihood that LLM_j generates f_{i_τ} based on the dialogue history h_τ, denoted as $G_{LLM_j}(h_\tau, f_{i_\tau})$. The overall dialogue quality score is derived by summing these log-likelihood scores, leading to the following formula:

$$QualityScore_{h_\tau} = \sum_{i=1}^{n} G_{LLM_j}(h_\tau, f_{i_\tau})$$

The formula described above is utilized to compute scores at the dialogue-level, assessing the quality of an entire conversation. However, it can be adapted for evaluation at the turn-level, where the objective is to assess the quality of a single response in the context of a given dialogue history. This adjustment can be accomplished by incorporating the response r into the formula, resulting in $G_{LLM_j}(h_\tau, r, f_{i_\tau})$.

The empirical effectiveness of utilizing conditional log-likelihood (as employed by FULL) compared to joint log-likelihood (as used by FED) is not well established (Kawamoto et al., 2023). To address this uncertainty, we implement the DIET metric in both settings. In the context of the dialogue history depicted in Figure 2 (an excerpt from the Strategic Conversations corpus (STAC) (Asher et al., 2016)), we evaluate the quality of the dialogue history h_{STAC}, including utterances such as "*Does anyone have wheat?*" from User A and "*Where is ash?*" from User B. For instance, using the conditional log-likelihood and follow-ups generated with the SELFMAP strategy (with model LLM_j as the backbone), we compute the following scores with LLM_j as the model evaluator: G_{LLM_j}("*That's not relevant to the game*" $\mid h_{STAC}), \ldots, G_{LLM_j}$("*I don't understand how that information relates to our game*" $\mid h_{STAC}$).

3.4 Negative Follow-Ups

Building on findings that negative follow-ups correlate more with human evaluations (De Bruyn et al., 2022), we focus solely on negative follow-ups. Using SELFMAP and CRAFT, we guide LLMs to generate these follow-ups by including a negative example in the meta-prompt. The final quality score is based on negative log-likelihood scores: higher scores indicate poor dialogue quality, while lower scores reflect better quality. The generated follow-ups are provided in Appendix B.

4 Discourse Structure Extraction Task

In line with previous research, we evaluate the DIET metric by its correlation with human judgments but argue that this alone may not comprehensively capture its effectiveness. Existing bench-

mark datasets for dialogue quality predominantly focus on open-domain conversations, which may not fully reflect DIET's domain-specific strengths. To further demonstrate its utility, we propose applying DIET to discourse analysis in dialogues, specifically for discourse structure extraction. This section provides an overview of the task and the application of DIET in addressing it.

Discourse analysis aims to extract a coherent structure from a document. One of the most commonly used framework — the Segmented Discourse Representation Theory (SDRT) framework (Asher and Lascarides, 2005) — structures a dialogue in a *Directed Acyclic Graph* (DAG), where clauses or sentences, known as *Elementary Discourse Units* (EDUs), are connected with rhetorical relations such as *Explanation* and *Elaboration*. In a recent work dealing with discourse structure extraction, Cimino et al. (2024) framed this task by decomposing a DAG structure into several sub-dialogues and proposed two algorithms for sub-dialogue construction (denoted as **DS-FLOW** and **DS-DP**). In order to identify the most coherent sub-dialogues, a metric that calculates the likelihood of a sequence of clauses is needed. Therefore, DIET can be used as the evaluation metric for forming sub-dialogues within both DS-FLOW and DS-DP[2].

5 Experimental Setup

Evaluation Tasks We assess DIET through two tasks: (i) Dialogue Quality Evaluation, where dialogue scores generated by automated methods are compared to human assessments; and (ii) Discourse Structure Extraction, which entails identifying interactions between EDUs to construct a DAG, as discussed above.

Models for Log-Likelihood and Follow-Up We use an open-source model to compute dialogue quality scores because proprietary models are limited to text-based prompts and do not allow output probability analysis. Moreover, since research suggests that LLMs fine-tuned on conversational data better capture dialogue quality aspects (Mehri and Eskénazi, 2020; De Bruyn et al., 2022) and we need an instruction-tuned model for log-likelihood analysis, we employ Vicuna-13b (Chiang et al., 2023) as the core model for DIET's evaluation step. In particular, Vicuna is based on the LLaMA model (Touvron et al., 2023) and trained on user-shared conversations, which is not the case for other possible alternatives such as Mistral (Jiang et al., 2023) and Pythia (Biderman et al., 2023). Consequently, Vicuna-13b is also used for follow-up generation in the SELFMAP strategy. In contrast, the CRAFT strategy uses distinct models for evaluation and generation, employing proprietary models for generating follow-ups, which have shown superior performance in dialogue tasks (Zhang et al., 2024). These models include Gemini 1.0 (Touvron et al., 2023), GPT-4 (OpenAI, 2023), and Claude 3[3].

Datasets We conduct experiments using three datasets for dialogue quality evaluation and two for discourse structure extraction. For dialogue quality evaluation, we use the **FED** dataset (Mehri and Eskénazi, 2020), consisting of 84 human-system conversations and 41 human-human conversations, annotated at both turn and dialogue levels with scores ranging from 0 to 4. We also use the **ConvAI2** dataset[4], which features 1157 open-domain chit-chat conversations about personal interests, annotated with scores from 1 to 5, after filtering out single-turn conversations. Lastly, we consider the **DSTC9** dataset (Gunasekara et al., 2020), which encompasses 2200 conversations between participants and knowledge-grounded response generation models, annotated with scores ranging from 1 to 5. For discourse structure extraction, we use the test set from the **STAC** dataset (Asher et al., 2016), which includes 109 documents detailing online interaction sessions of the game "The Settlers of Catan". We also use the cleaned version of the **Molweni** dataset (Li et al., 2020), **Molweni-clean** (Li et al., 2024a), containing 50 documents focused on technical discussions about Ubuntu operating system. Detailed dataset statistics are in Table 1.

Evaluation Metrics To evaluate the correlation between the predicted quality scores and human annotations, we use the Spearman rank correlation coefficient. Conversely, to assess the effectiveness of the DIET metric in sub-dialogue evaluation for the discourse structure extraction task, we report the micro-F_1 score for the generated structures.

Compared Methods For both tasks, we compare the performance of the DIET metric (using both

[2]Appendix C offers a comparative analysis of the performance of the perplexity metric, as used in Cimino et al. (2024) for sub-dialogue evaluation, in relation to the DIET, FULL, and FED metrics.

[3]https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku

[4]https://github.com/DeepPavlovAdmin/convai/tree/master/

Task	Dataset	Size	Diag Len	Utt Len
Discourse Structure Extraction	STAC	109	175	16
	Molweni-clean	50	447	52
Dialogue Quality Evaluation	FED dataset	125	168	11
	ConvAI2	1157	573	33
	DSTC9	2200	318	11

Table 1: Statistics on the number of dialogues, as well as the average dialogue and utterance lengths across the datasets.

meta-prompting strategies) against the FED and FULL metrics. For the FED and FULL metrics, we use the follow-ups specified in their respective studies but adapt them to evaluate human-to-human dialogues rather than just human-system interactions. To tailor these metrics to our context, we analyze the follow-ups and propose an alternative version of the FED metric that excludes certain dialogue qualities, specifically *diverse*, *likeable*, *flexible*, *informative*, and *inquisitive*. For instance, we omit the dialogue quality categorized as "*likeable*", which includes follow-ups such as "*You're super polite and fun to talk to*", as it is less relevant in settings like the Catan game, where the focus is on strategic interaction rather than conversational engagement. By customizing these metrics, we aim to accurately assess dialogue quality in our specific use case, ensuring alignment with the dynamics of human-to-human interaction (Kawamoto et al., 2023). In contrast, we keep all five follow-ups used by FULL, as they are broadly applicable. The list of follow-ups for the FED and FULL metrics is detailed in Appendix D.

6 Results

We evaluate DIET by examining two key aspects: the type of follow-ups used (domain-specific versus general-purpose utterances) and the method for computing quality scores (joint log-likelihood versus conditional log-likelihood).

Performance in Dialogue Quality Evaluation Benchmark datasets for dialogue quality evaluation are open-domain, complicating the creation of a specific D_τ for the meta-prompt. Therefore, we use general-purpose follow-ups. Table 2 shows the Spearman coefficients for DIET, FED, and FULL metrics across the FED, ConvAI2, and DSTC9 datasets, indicating whether each metric uses joint (JLL) or conditional (CLL) log-likelihood. FED_{sub} uses a subset of dialogue qualities from the FED metric, while FED_{com} includes all 18 qualities of the FED metric. Results for FED_{com} and FULL on the FED dataset are from De Bruyn et al. (2022), with other results replicated.

Examining meta-prompting strategies reveals compositionality does not affect performance when using generic follow-ups. Using consistent log-likelihood methods for calculating quality scores, DIET yields similar results across both meta-prompting strategies. This consistency is due to the general nature of follow-ups encountered by large models like Vicuna-13b during training.

Table 2 shows that with joint log-likelihood, DIET consistently outperforms FED and FULL in two out of three datasets in dialogue-level evaluations, regardless of the meta-prompting strategy. With conditional log-likelihood, DIET exceeds FED_{sub} and FED_{com} in turn-level evaluations but falls short of FULL. DIET's lower performance on the FED dataset likely stems from the follow-up selection bias in FED and FULL for this dataset (Kawamoto et al., 2023). Averaging Spearman coefficients across all dialogue-level evaluations, DIET with joint log-likelihood scores 0.31, exceeding FULL, FED_{sub}, and FED_{com} at 0.30, 0.04, and 0.12, respectively. Notably, DIET's follow-ups were generated automatically without human intervention or annotated data analysis.

Using conditional log-likelihood improves DIET's performance on the FED dataset at both turn and dialogue levels, though improvements vary across datasets. In contrast, joint log-likelihood offers more stable and reliable performance by considering the entire dialogue and follow-ups, helping the model capture context for consistent scores.

Performance in Discourse Structure Extraction Unlike dialogue quality evaluation, discourse structure extraction enables the generation of domain-specific follow-ups for the STAC and Molweni-clean datasets. Table 3 shows the micro-F_1 scores obtained using DIET, FED, and FULL as evaluation metrics in DS-FLOW and DS-DP.

When assessing dialogues within a specific domain, conditional log-likelihood often improves outcomes. Although joint log-likelihood occasionally yields higher micro-F_1 score — such as with $\text{DIET}_{SELFMAP-JLL}$ and $\text{DIET}_{CRAFT-JLL}$ using DS-DP on the Molweni-clean dataset — conditional log-likelihood generally performs better on average. For instance, $\text{DIET}_{SELFMAP}$ scores 70.6% (53.9%, respectively) with joint

Dialogue Quality Evaluation	FED dataset		ConvAI2	DSTC9
Metric	**Turn-Level**	**Dialogue-Level**	**Dialogue-Level**	**Dialogue-Level**
FED_{sub}	0.09	0.15	−0.13	0.10
FED_{com}	0.09	0.32	−0.06	0.10
FULL	0.51	0.69	0.12	0.10
$\text{DIET}_{SELFMAP-JLL}$	0.08	0.25	**0.57**	**0.11**
$\text{DIET}_{SELFMAP-CLL}$	0.44	0.47	0.05	0.07
$\text{DIET}_{CRAFT-JLL}$	0.08	0.25	**0.57**	**0.11**
$\text{DIET}_{CRAFT-CLL}$	**0.49**	**0.53**	0.08	0.05

Table 2: Spearman correlations with human judgment. Bold values indicate the highest DIET scores. Underlined scores show the best performance for each setting. JLL: Joint Log-Likelihood. CLL: Conditional Log-Likelihood.

Discourse Structure Extraction	STAC		Molweni-clean	
Metric	**DS-FLOW**	**DS-DP**	**DS-FLOW**	**DS-DP**
FED_{sub}	37.2	33.0	39.1	31.6
FED_{com}	39.4	34.6	44.1	35.0
FULL	55.3	40.8	68.8	41.4
$\text{DIET}_{SELFMAP-JLL}$	54.5	53.3	64.2	**76.9**
$\text{DIET}_{SELFMAP-CLL}$	**57.2**	55.8	76.5	67.3
$\text{DIET}_{CRAFT-JLL}$	54.1	52.2	62.8	**76.9**
$\text{DIET}_{CRAFT-CLL}$	**57.2**	**56.5**	**76.8**	67.9

Table 3: Micro-F_1 scores for the discourse structure extraction task on STAC and Molweni-clean datasets using DIET, FULL, and FED for sub-dialogue evaluation. Highest scores are in bold. JLL: Joint Log-Likelihood. CLL: Conditional Log-Likelihood. DS-FLOW and DS-DP are proposed in Cimino et al. (2024).

log-likelihood versus 71.9% (56.5%, respectively) with conditional log-likelihood on the Molweni-clean (STAC, respectively) dataset. Similarly, DIET_{CRAFT} achieves 69.9% (53.2%, respectively) with joint log-likelihood and 72.4% (56.9%, respectively) with conditional log-likelihood. Additionally, conditional log-likelihood reduces reliance on the model generating follow-ups. Specifically, $\text{DIET}_{CRAFT-CLL}$ often outperforms or matches $\text{DIET}_{SELFMAP-CLL}$, likely due to the robustness of conditional log-likelihood, as it is less affected by the specific follow-up generation model, provided the follow-ups are relevant and contextually appropriate. Consequently, using diverse models for follow-up generation introduces variability and mitigates biases, leading to more effective evaluations of domain-specific dialogues.

When using joint log-likelihood, compositionality's effect becomes clear. Specifically, an analysis of both algorithms on the STAC and Molweni-clean datasets shows that follow-ups generated by Vicuna-13b ($\text{DIET}_{SELFMAP-JLL}$) often outperform those from Gemini 1.0, GPT-4, and Claude 3 ($\text{DIET}_{CRAFT-JLL}$). This is likely because domain-specific follow-ups reduce the likelihood that a model has seen the CRAFT-generated utterances before. As previously highlighted, follow-ups familiar to the model during joint log-likelihood analysis boost performance.

Overall, DIET significantly outperforms both FED and FULL in domain-specific dialogue evaluation, highlighting the limitations of using general-purpose follow-ups with FED and FULL across different domains. The results also indicate that some dialogue qualities, though seemingly irrelevant for assessing human-to-human interactions, are actually valuable. Notably, FED_{com} performed better than FED_{sub} in both datasets and algorithms.

7 Conclusion

This paper introduces DIET, a novel unsupervised and reference-free metric for dialogue evaluation. DIET uses meta-prompting with LLMs to automatically generate domain-specific follow-ups, analyzed through log-likelihood to assess dialogue quality. We propose two meta-prompting strategies: SELFMAP, where a single model generates and evaluates follow-ups, and CRAFT, which uses an ensemble of models. Intrinsic experiments show that DIET outperforms existing reference-free metrics in aligning with human judgments. In extrinsic experiments, DIET excels in real-world discourse parsing tasks, generating contextually relevant, domain-specific follow-ups. These results confirm DIET's effectiveness as a robust tool for dialogue evaluation across domains.

8 Limitations

While DIET shows strong performance in assessing dialogue quality, its reliance on large models leads to increased computational time when calculating probabilities for longer dialogues. Similar to FED and FULL, DIET requires the computation of multiple log-likelihood values for evaluation, significantly raising the time complexity as the dialogue length grows. This challenge is especially evident in extended interactions, where the computational overhead may impede real-time processing. As a result, DIET's reliance on extensive model parameters and multi-step probability calculations presents scalability issues, particularly in resource-constrained environments. Therefore, optimizing the metric's efficiency without sacrificing assessment accuracy is a key focus of ongoing research and development.

The CRAFT strategy provides a means to evaluate whether compositionality affects dialogue evaluation. However, it does not guarantee that the follow-ups generated by an ensemble of models $LLM_1, \ldots, LLM_m$ have not been previously encountered by the model LLM_j used for dialogue evaluation, where $j \neq 1, \ldots, m$. Nevertheless, the likelihood of LLM_j encountering these utterances is significantly lower than that of those generated by LLM_j itself. This issue will be addressed in future work, incorporating methods to verify if an example exists within the corpus used for pre-training an LLM (Shi et al., 2023). Additionally, refining CRAFT to better select and combine follow-ups from an ensemble could mitigate biases and enhance performance.

Current strategies, including FED and FULL, may produce less granular follow-ups. Specifically, employing a uniform set of follow-up utterances across all dialogues can diminish the specificity and relevance of these responses. This generalized approach risks neglecting the unique contexts and nuances that characterize individual conversations. Therefore, instance-aware prompt learning (Jin et al., 2023) presents an opportunity to customize follow-ups for each dialogue, rather than depending on a single, standardized set applicable to all interactions. By utilizing instance-aware techniques, we can tailor the follow-ups to reflect the distinct attributes of each interaction, resulting in more contextually appropriate and effective responses.

9 Ethical Considerations

This research introduces the DIET metric, an unsupervised and reference-free approach for dialogue evaluation, utilizing LLMs to automatically generate follow-up utterances. As with any AI system, the ethical implications of deploying LLMs in unsupervised evaluation must be carefully considered.

Firstly, the reliance on LLMs raises concerns regarding potential biases embedded in the training data. LLMs may generate follow-up utterances that inadvertently reflect stereotypes or cultural biases, which could skew the evaluation of dialogues, especially in cross-cultural or sensitive contexts. To mitigate this, we employ multiple models in the CRAFT strategy, aiming to balance individual model biases and achieve more diverse and fair outputs. However, further steps, such as bias auditing and mitigation techniques, are essential to ensure that dialogue evaluations are not unfairly influenced by these biases.

In light of these concerns, we have meticulously curated the dialogue corpora used in this study to minimize biases, hate speech, and inappropriate language. This involves employing human-annotated datasets and professionally curated resources. Additionally, we take into account the privacy of dialogue partners by replacing names with generic user tokens within the selected datasets.

Secondly, this work involves generating synthetic data (i.e., follow-up utterances) using LLMs. While this data is used for the purpose of evaluating dialogue quality, there remains a risk that the generated content could unintentionally produce harmful, inappropriate, or misleading responses, particularly when models are used without stringent filtering mechanisms. We have carefully designed meta-prompting strategies to reduce such risks, but ethical scrutiny of the output should continue, especially if such methods are deployed in real-world applications.

Finally, the use of LLMs introduces environmental concerns due to the computational resources required for training and deploying such models. While the presented methodology minimizes the need for large-scale supervised datasets, the computation necessary for generating and evaluating follow-ups still involves significant energy consumption. Future work should consider more energy-efficient models or techniques to mitigate the environmental impact of this approach.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Gaetano Cimino, Chuyuan Li, Giuseppe Carenini, and Vincenzo Deufemia. 2024. Coherence-based dialogue discourse structure extraction using open-source large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–316.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2022. Open-domain dialog evaluation using follow-ups likelihood. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 496–504. International Committee on Computational Linguistics.

Adrian de Wynter, Xun Wang, Qilong Gu, and Si-Qing Chen. 2023. On meta-prompting. *CoRR*, abs/2312.06562.

Suvodip Dey and Maunendra Sankar Desarkar. 2023. Dial-m: A masking-based framework for dialogue evaluation. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2023, Prague, Czechia, September 11 - 15, 2023*, pages 77–84. Association for Computational Linguistics.

Maxine Eskénazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. 2019. Beyond turing: Intelligent agents centered on the user. *CoRR*, abs/1901.06613.

Sarah E. Finch, Ellie S. Paek, and Jinho D. Choi. 2023. Leveraging large language models for automated dialogue analysis. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2023, Prague, Czechia, September 11 - 15, 2023*, pages 202–215. Association for Computational Linguistics.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 445–450. The Association for Computer Linguistics.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *CoRR*, abs/1904.10635.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: dialogue coherence evaluation using amr-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 771–785. Association for Computational Linguistics.

R. Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David R. Traum, Maxine Eskénazi, Ahmad Beirami, Eunjoon Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. Overview of the ninth dialog system technology challenge: DSTC9. *CoRR*, abs/2011.06486.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 2019.

Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 379–391. Association for Computational Linguistics.

Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. 2022. Metaprompting: Learning to learn better prompts. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3251–3262. International Committee on Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9230–9240. Association for Computational Linguistics.

Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Galstyan, and Anoop Kumar. 2024. Leveraging llms for dialogue quality measurement. *CoRR*, abs/2406.17304.

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

Feihu Jin, Jinliang Lu, Jiajun Zhang, and Chengqing Zong. 2023. Instance-aware prompt learning for language understanding and generation. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(7):199:1–199:18.

Toshiki Kawamoto, Yuki Okano, Takato Yamazaki, Toshinori Sato, Kotaro Funakoshi, and Manabu Okumura. 2023. A follow-up study on evaluation metrics using follow-up utterances. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, PACLIC 2023, The Hong Kong Polytechnic University, Hong Kong, SAR, China, 2-4 December 2023*, pages 552–558. Association for Computational Linguistics.

Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. 2024a. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2642–2652. International Committee on Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Xue Li, Jia Su, Yang Yang, Zipeng Gao, Xinyu Duan, and Yi Guan. 2024b. Dialogues are not just text: Modeling cognition for dialogue coherence evaluation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18573–18581. AAAI Press.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *CoRR*, abs/2305.13711.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1116–1126. Association for Computational Linguistics.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskénazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David R. Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the NSF future directions workshop on automatic evaluation of dialog: Research directions and challenges. *CoRR*, abs/2203.10012.

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

John Mendonça, Alon Lavie, and Isabel Trancoso. 2022. Qualityadapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022, Edinburgh, UK, 07-09 September 2022*, pages 83–90. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

ChaeHun Park, Eugene Jang, Wonsuk Yang, and Jong Park. 2021. Generating negative samples by manipulating golden responses for unsupervised learning of a response evaluation model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1525–1534. Association for Computational Linguistics.

ChaeHun Park, Seungil Chad Lee, Daniel Rim, and Jaegul Choo. 2023. Density: Open-domain dialogue evaluation metric using density estimation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14222–14236. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *CoRR*, abs/2310.16789.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 722–729. AAAI Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. *CoRR*, abs/2308.05342.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *CoRR*, abs/2106.03706.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. 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.

Chen Zhang, Luis Fernando D'Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2020a. Deep AM-FM: toolkit for automatic dialogue evaluation. In *Conversational Dialogue Systems for the Next Decade - 11th International Workshop on Spoken*

Dialogue Systems, IWSDS 2020, Madrid, Spain, 21-23 September, 2020, volume 704 of *Lecture Notes in Electrical Engineering*, pages 53–69. Springer.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19515–19524. AAAI Press.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. 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 1815–1825.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Zhenyu Zhang, Tao Guo, and Meng Chen. 2021b. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management, CIKM '21, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3647–3651. ACM.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

A Meta-Prompts

Considering that STAC and Molweni-clean datasets are domain-specific, we have formally designed two customized descriptions, D_{STAC} and $D_{Molweni}$, respectively, as illustrated in Figure 3. In contrast, for the FED, ConvAI2, and DSTC9 datasets, which comprise chit-chat conversations, we edit the final part of the meta-prompt by instructing LLMs to "*propose a set of potential follow-up utterances that could be employed to evaluate responses in general conversations*", thereby creating general-purpose follow-ups.

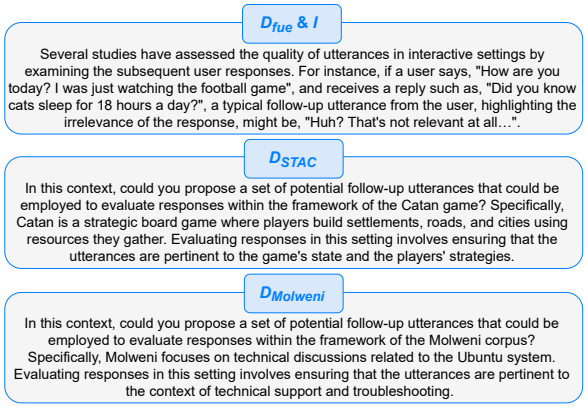

Figure 3: Meta-prompt components designed to generate follow-ups in dialogue evaluation using the DIET metric.

B Generated Follow-Ups

Figures 4-6 show follow-ups generated using the SELFMAP strategy, with Vicuna-13b as the core model. The initial two sets of follow-ups are domain-specific, utilizing the D_{STAC} and $D_{Molweni}$ descriptions, respectively. In contrast, the final set is generated based on general instruction, resulting in general-purpose follow-ups. Figures 7-9, on the other hand, illustrate follow-ups generated through the CRAFT strategy, employing the Gemini 1.0, Claude 3, and GPT-4 models. To align with the average number of follow-ups produced by the Vicuna-13b model, the Gemini 1.0, Claude 3, and GPT-4 models were each prompted to generate three follow-ups per context, resulting in nine follow-ups per context.

C Comparison with Perplexity

In this section, we compare the performance of the perplexity metric, as adopted in Cimino et al. (2024), against the DIET, FULL, and FED metrics

SELFMAP (STAC)

. *That's not relevant to the game. We're currently building our settlements and gathering resources.*
. *I don't think that information is helpful. We need to know what cards are available for us to trade.*
. *Can you give us some advice on how to improve our strategy? That would be more helpful than sharing random facts.*
. *I don't understand how that information relates to our game. Can you please provide something more relevant?*
. *Thanks for sharing, but we need to focus on our current game state. Can you give us some insights on how to proceed?*
. *That's interesting, but it doesn't help us with our current decision. Can you provide some guidance on our next move?*
. *I don't think that information is useful for our current situation. Can you give us some hints on how to proceed with our game?*
. *That's not related to our game. We need help with our current strategy, not general knowledge.*

Figure 4: Follow-ups generated using the meta-prompt, incorporating the D_{STAC} description, with the Vicuna-13b model for discourse structure extraction on the STAC dataset.

SELFMAP (Molweni)

. *That's not helpful. Can you please provide more specific guidance on how to resolve the issue?*
. *I don't understand how that relates to my problem. Can you please rephrase your response or provide more relevant information?*
. *Can you please provide a step-by-step solution or troubleshooting guide for this issue?*
. *I'm still having trouble with my system. Can you please provide additional assistance or escalate this issue to a higher level of support?*
. *Can you please confirm if the solution you provided is compatible with the Ubuntu version I'm using?*
. *That response doesn't seem to address my question. Can you please provide more context or clarify your answer?*
. *Can you please provide more detailed information about the technical aspects of the issue I'm experiencing?*
. *Can you please offer alternative solutions or further assistance?*
. *Can you please provide more examples or explain the concept in simpler terms?*
. *I'm not sure I understand the technical jargon you used. Can you please rephrase your response using more accessible language?*

Figure 5: Follow-ups generated using the meta-prompt, incorporating the $D_{Molweni}$ description, with the Vicuna-13b model for discourse structure extraction on the Molweni-clean dataset.

SELFMAP (General)

. *Huh? That's not relevant at all...*
. *Why did you say that?*
. *I wasn't expecting that response...*
. *Can you repeat that?*
. *Can you give me more information on that?*
. *I don't understand...*
. *That's interesting, but it's not what I was asking for...*
. *Can you give me a more detailed explanation?*
. *I think you misunderstood my question...*
. *That's not what I was expecting to hear...*

Figure 6: Follow-ups generated using the meta-prompt with the Vicuna-13b model for dialogue quality evaluation.

in both dialogue quality evaluation and discourse structure extraction tasks.

From Table 4, we can observe that perplexity fails to produce scores that correlate with human judgments for dialogue quality evaluation. It consistently underperforms compared to the FED_{sub} metric, which itself yielded the poorest results among the metrics employing follow-ups for di-

CRAFT (STAC)

Gemini

. *That's interesting, but I'm not sure how it applies to the current situation.*
. *Huh? That's not relevant at all...*
. *I see what you're saying, but I think there might be a better option.*

. *That move doesn't align with the trading opportunities available.*
. *That strategy doesn't make sense given the current board state.*
. *I don't think that's the optimal play right now.*

. *That doesn't seem relevant to the current situation.*
. *That advice doesn't match the current game situation.*
. *That's not aligned with my current objective.*

Figure 7: Follow-ups generated using the meta-prompt, incorporating the D_{STAC} description, with the Gemini 1.0, Claude 3, and GPT-4 models for discourse structure extraction on the STAC dataset.

CRAFT (Molweni)

Gemini

. *That's not relevant to the problem I'm describing.*
. *Can you try again and provide a more relevant answer?*
. *I'm not sure how that information helps me troubleshoot the issue.*

. *I'm afraid I don't understand how your response helps with my technical problem.*
. *Could you please clarify how your response relates to my Ubuntu issue?*
. *I appreciate the information, but it's not directly addressing the Ubuntu-related problem I described.*

. *That doesn't seem to help with my Ubuntu issue…*
. *I need more relevant details about Ubuntu.*
. *I need help specifically with Ubuntu, not something else.*

Figure 8: Follow-ups generated using the meta-prompt, incorporating the $D_{Molweni}$ description, with the Gemini 1.0, Claude 3, and GPT-4 models for discourse structure extraction on the Molweni-clean dataset.

CRAFT (General)

Gemini

. *I'm not sure how your response relates to my previous statement.*
. *That doesn't seem to follow logically from what we were talking about.*
. *That's not very helpful.*

. *I'm not following how your response relates to our discussion.*
. *Your answer doesn't seem to address my question/comment.*
. *That seems like a non-sequitur. Can we get back on track?*

. *That doesn't make sense.*
. *Could you clarify what you just said?*
. *Let's talk about something else.*

Figure 9: Follow-ups generated using the meta-prompt with the Gemini 1.0, Claude 3, and GPT-4 models for dialogue quality evaluation.

alogue evaluation. Consequently, DIET demonstrates a significant advantage over perplexity in terms of correlation with human judgments for dialogue quality. Conversely, Table 5 shows that perplexity surpasses DIET in the discourse structure extraction task, particularly over the STAC corpus using both DS-FLOW and DS-DP. This suggests that while perplexity is ineffective in capturing aspects of dialogue quality that align with human perception, it remains a reliable metric for assessing the coherence of sub-dialogues within the discourse structure extraction task. This effectiveness may be attributed to perplexity's ability to model local coherence and flow within smaller discourse units, which is crucial for constructing coherent sub-dialogues. In contrast, DIET excels in evaluating overall dialogue quality and coherence, providing a better alignment with human judgments at the conversation level.

D List of FED and FULL Follow-Ups

Table 6 presents the list of follow-ups utilized by the FULL, FED_{com}, and FED_{sub} metrics. It categorizes each follow-up by its specific metric, category, level, and type, distinguishing between positive and negative utterances.

Dialogue Quality Evaluation	FED dataset		ConvAI2	DSTC9
Metric	**Turn-Level**	**Dialogue-Level**	**Dialogue-Level**	**Dialogue-Level**
FED_{sub}	0.09	0.15	−0.13	0.10
FED_{com}	0.09	0.32	−0.06	0.10
FULL	<u>0.51</u>	<u>0.69</u>	0.12	0.10
Perplexity	−0.17	−0.099	−0.34	−0.18
$\text{DIET}_{SELFMAP-JLL}$	0.08	0.25	**<u>0.57</u>**	**<u>0.11</u>**
$\text{DIET}_{SELFMAP-CLL}$	0.44	0.47	0.05	0.07
$\text{DIET}_{CRAFT-JLL}$	0.08	0.25	**<u>0.57</u>**	**<u>0.11</u>**
$\text{DIET}_{CRAFT-CLL}$	**0.49**	**0.53**	0.08	0.05

Table 4: Spearman correlations with human judgment. Bold values indicate the highest DIET scores. Underlined scores represent the best performance for each setting. JLL: Joint Log-Likelihood. CLL: Conditional Log-Likelihood.

Discourse Structure Extraction	STAC		Molweni-clean	
Metric	**DS-FLOW**	**DS-DP**	**DS-FLOW**	**DS-DP**
FED_{sub}	37.2	33.0	39.1	31.6
FED_{com}	39.4	34.6	44.1	35.0
FULL	55.3	40.8	68.8	41.4
Perplexity	<u>58.1</u>	<u>57.3</u>	72.9	74.7
$\text{DIET}_{SELFMAP-JLL}$	54.5	53.3	64.2	**<u>76.9</u>**
$\text{DIET}_{SELFMAP-CLL}$	**57.2**	55.8	76.5	67.3
$\text{DIET}_{CRAFT-JLL}$	54.1	52.2	62.8	**<u>76.9</u>**
$\text{DIET}_{CRAFT-CLL}$	**57.2**	**56.5**	**<u>76.8</u>**	67.9

Table 5: Micro-F_1 scores for the discourse structure extraction task on STAC and Molweni-clean datasets using DIET, FULL, FED, and Perplexity for sub-dialogue evaluation. Bold values indicate the highest DIET scores. Underlined scores represent the best performance for each setting. JLL: Joint Log-Likelihood. CLL: Conditional Log-Likelihood. DS-FLOW and DS-DP are proposed in Cimino et al. (2024).

Metric(s)	Follow-Up	Category	Level	Type
FULL,FED$_{com}$,FED$_{sub}$	Not really relevant here.	Specific	Turn	Negative
FULL,FED$_{com}$,FED$_{sub}$	You're really confusing.	Error recovery	Dialogue	Negative
FULL,FED$_{com}$	You're really boring.	Informative	Dialogue	Negative
FULL,FED$_{com}$	You don't seem interested.	Inquisitive	Dialogue	Negative
FULL,FED$_{com}$,FED$_{sub}$	What are you trying to say?	Understanding	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	I don't understand what you're saying.	Correct	Turn	Negative
FED$_{com}$,FED$_{sub}$	That's not really relevant here.	Specific	Turn	Negative
FED$_{com}$,FED$_{sub}$	You are so confusing.	Coherent	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	That's not very interesting.	Interesting	Turn	Negative
FED$_{com}$,FED$_{sub}$	That was a really boring response.	Interesting	Turn	Negative
FED$_{com}$,FED$_{sub}$	I am so confused right now.	Error recovery	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	I'm so confused!	Understandable	Turn	Negative
FED$_{com}$,FED$_{sub}$	I don't really care. That's pretty boring.	Engaging	Turn	Negative
FED$_{com}$,FED$_{sub}$	I want to talk about something else.	Engaging	Turn	Negative
FED$_{com}$,FED$_{sub}$	That's not even related to what I said.	Relevant	Turn	Negative
FED$_{com}$,FED$_{sub}$	I am so confused right now!	Correct	Turn	Negative
FED$_{com}$,FED$_{sub}$	That makes no sense!	Semantically appropriate	Turn	Negative
FED$_{com}$,FED$_{sub}$	I don't understand at all!	Understandable	Turn	Negative
FED$_{com}$,FED$_{sub}$	That's really boring.	Interesting	Turn	Negative
FED$_{com}$	I don't like you.	Likeable	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	I'm so confused right now!	Fluent	Turn	Negative
FED$_{com}$,FED$_{sub}$	Don't change the topic!	Relevant	Turn	Negative
FED$_{com}$,FED$_{sub}$	You're not understanding me!	Correct	Turn	Negative
FED$_{com}$,FED$_{sub}$	That's a very generic response.	Specific	Turn	Negative
FED$_{com}$	You don't really know much.	Informative	Dialogue	Negative
FED$_{com}$	You're not very nice.	Likeable	Dialogue	Negative
FED$_{com}$	You're not very fun to talk to.	Likeable	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	Is that real English?	Fluent	Turn	Negative
FED$_{com}$	That's a lot of questions!	Inquisitive	Dialogue	Positive
FED$_{com}$	Why are you repeating yourself?	Diverse	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	You're making no sense at all.	Coherent	Dialogue	Negative
FED$_{com}$	You ask a lot of questions!	Inquisitive	Dialogue	Positive
FED$_{com}$,FED$_{sub}$	Let's change the topic.	Engaging	Turn	Negative
FED$_{com}$	You don't ask many questions.	Inquisitive	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	Why are you changing the topic?	Relevant	Turn	Negative
FED$_{com}$	Stop saying the same thing repeatedly.	Diverse	Dialogue	Negative
FED$_{com}$	Do you know how to talk about something else?	Flexible	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	You're changing the topic so much!	Coherent	Dialogue	Negative
FED$_{com}$	You know a lot of facts!	Informative	Dialogue	Positive
FED$_{com}$,FED$_{sub}$	Tell me more!	Engaging	Turn	Positive
FED$_{com}$	I like you!	Likeable	Dialogue	Positive
FED$_{com}$	Wow that's a lot of information.	Informative	Dialogue	Positive
FED$_{com}$,FED$_{sub}$	Stop changing the topic so much.	Depth	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	What does that even mean?	Understandable	Turn	Negative
FED$_{com}$	I don't want to talk about that!	Flexible	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	That's not what you said earlier!	Consistent	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	You have a good point.	Semantically appropriate	Turn	Positive
FED$_{com}$,FED$_{sub}$	I see, that's interesting.	Specific	Turn	Positive
FED$_{com}$,FED$_{sub}$	Stop contradicting yourself!	Consistent	Dialogue	Negative
FED$_{com}$	You're very easy to talk to!	Flexible	Dialogue	Positive
FED$_{com}$	Stop repeating yourself!	Diverse	Dialogue	Negative
FED$_{com}$,FED$_{sub}$	That's good to know. Cool!	Specific	Turn	Positive
FED$_{com}$,FED$_{sub}$	That's a good point.	Specific	Turn	Positive
FED$_{com}$	Wow you can talk about a lot of things!	Flexible	Dialogue	Positive
FED$_{com}$,FED$_{sub}$	I'm really interested in learning more about this.	Engaging	Turn	Positive
FED$_{com}$,FED$_{sub}$	That makes sense!	Semantically appropriate	Turn	Positive
FED$_{com}$	Thanks for all the information!	Informative	Dialogue	Positive
FED$_{com}$	You're super polite and fun to talk to.	Likeable	Dialogue	Positive
FED$_{com}$,FED$_{sub}$	Wow that is really interesting.	Interesting	Turn	Positive
FED$_{com}$,FED$_{sub}$	That's really interesting!	Interesting	Turn	Positive
FED$_{com}$	Great talking to you.	Likeable	Dialogue	Positive
FED$_{com}$,FED$_{sub}$	Cool! That sounds super interesting.	Interesting	Turn	Positive
FED$_{com}$,FED$_{sub}$	Wow! That's really cool!	Engaging	Turn	Positive

Table 6: List of follow-ups employed by FULL, FED$_{com}$, and FED$_{sub}$, categorized by fine-grained attribute, level (turn or dialogue), and polarity (positive or negative).