Meta-Prompting Follow-Ups for Unsupervised Dialogue Evaluation using Open-Source Large Language Models

Gaetano Cimino[δ], **Chuyuan Li**[δ], **Giuseppe Carenini**[δ], **Vincenzo Deufemia**[γ]
[δ]University of British Columbia, Canada
[γ]University of Salerno, Italy
{gaetano.cimino, chuyuan.li}@ubc.ca, carenini@cs.ubc.ca, deufemia@unisa.it

Abstract

Automatically evaluating dialogue quality remains a major challenge due to the complexity and contextual variability of human interactions. This paper introduces DIET, a novel unsupervised, reference-free metric that uses follow-up utterances to assess dialogue quality. Unlike existing reference-free metrics, which rely on follow-ups derived from annotated data and apply a uniform set of utterances across all dialogues, DIET generates follow-ups using open-source Large Language Models (LLMs) and refines them through a selection process. Two strategies are explored: SELFMAP, where generation and evaluation are performed by the same model to ensure internal coherence, and CRAFT, where multiple models collaborate to generate diverse and complementary follow-ups, enhancing robustness and reducing model bias. Dialogue quality is measured via the likelihood of an LLM continuing the dialogue from selected follow-ups. Experiments show DIET better correlates with human judgments than existing reference-free metrics across multiple meta-evaluation datasets.

Keywords: Dialogue Evaluation, Reference-Free Metrics, Meta-Prompting

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., 2020a; Mehri et al., 2022). 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 (Liu et al., 2016; Yeh et al., 2021). Furthermore, few annotated dialogues hamper the development of effective supervised metrics (Zhang et al., 2022), underscoring the need for accurate unsupervised alternatives.

Follow-up utterances have proven to be an effective means of evaluating dialogue quality, serving as implicit feedback from users (Eskénazi et al., 2019). Leveraging this approach, metrics such as FED (Mehri and Eskénazi, 2020) and FULL (De Bruyn et al., 2022) assess dialogue quality by estimating the likelihood that language models assign to predefined follow-ups derived from annotated data. However, since the appropriateness of a follow-up is context-dependent, reliance on a fixed set of follow-ups may limit these metrics' ability to capture the complexity of diverse dialogue scenarios. Furthermore, not all follow-ups reliably indicate dialogue quality, as some are more contextually plausible than others. For example, in the sample dialogue shown in Figure 1, the final question posed by the first speaker, "*Do more people live in the USA or Australia?*", is misinterpreted by the second speaker as a Yes/No question, leading to the response "*Not that much more, no*". Consequently, the follow-up "*That's an interesting fact, but it doesn't answer my query*" is implausible, as it wrongly implies the second speaker introduced an off-topic fact. In contrast, the follow-up "*I think you misunderstood my question*" more accurately captures the issue by addressing the misinterpretation of the first speaker's intent. Thus, selecting contextually relevant follow-ups improves evaluation accuracy by reducing irrelevant responses, mitigating bias, and better reflecting model performance.

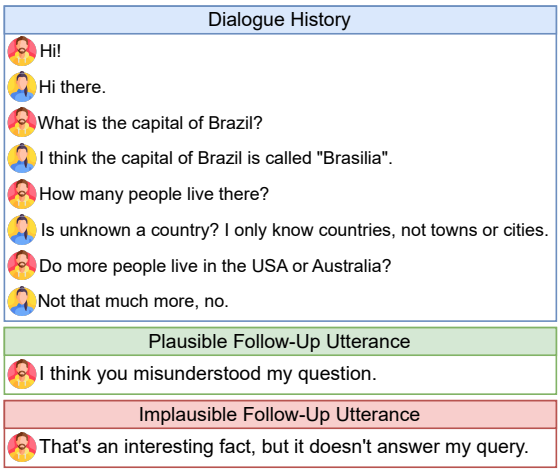

Figure 1: A dialogue history from the FED dataset (Mehri and Eskénazi, 2020) with plausible and implausible follow-up utterances.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2812–2824
11–16 May 2026.

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 follow-ups generated through meta-prompting strategies with open-source Large Language Models (LLMs) (Cimino and Deufemia, 2025). Meta-prompting aims to create prompts that help LLMs generate contextually relevant and insightful follow-ups tailored to dialogue evaluation. Two meta-prompting strategies are proposed: Self-Referential Follow-Up Meta-Prompting (SELF-MAP), 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. Moreover, DIET employs a methodology to select dialogue-specific follow-ups, filtering out irrelevant ones that may affect the accuracy of quality assessments.

We assess DIET across multiple meta-evaluation datasets, demonstrating its superiority over prior metrics when employing the CRAFT strategy in conjunction with conditional log-likelihood computation. Further analysis confirms its reliability in evaluating open-domain conversations.

The main contributions of this paper include: (i) DIET, a new unsupervised, reference-free metric for dialogue quality evaluation; (ii) two LLM-based meta-prompting strategies to auto-generate follow-ups; (iii) a process to select follow-ups that best align with each dialogue's context; and (iv) an evaluation showing DIET's improved alignment with human judgments over state-of-the-art metrics.

2. Related Work

Automatic 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.

2.1. Reference-Based Metrics

These metrics compare generated responses to reference utterances via word-overlap methods (e.g., BLEU (Papineni et al., 2002)) or learning-based techniques (e.g., BLEURT (Sellam et al., 2020)). Such metrics often misalign with human judgment, primarily due to the one-to-many nature of dialogue (Liu et al., 2016; Yeh et al., 2021). Furthermore, their practical use is limited: they are unsuitable for online evaluation, where reference responses are unavailable and require costly human annotation.

[1]The name underscores a key advantage over FED and FULL: no annotated data is required for follow-ups.

2.2. Reference-Free Metrics

These metrics assess dialogue quality without the need for human-defined references. Most of the metrics proposed in the literature rely on the comparison of positive and negative samples. For instance, DynaEval (Zhang et al., 2021) uses graph convolutional networks to assess dialogues modeled as graphs, trained on negative samples created by altering utterances or speaker order. Similarly, QualityAdapt (Mendonça et al., 2022) uses the Adapter paradigm (Houlsby et al., 2019) to train individual adapters on dialogue subqualities, generating negatives by word reordering, dropping, and repetition. Conversely, DEnsity (Park et al., 2023) scores responses via density estimation on the feature space derived from a response selection model trained to distinguish correct responses from randomly sampled negative examples. However, these methods rely on artificial negative samples, missing subtle incoherence in human conversations (Ghazarian et al., 2022).

Recent work has explored LLMs for dialogue evaluation. Park et al. (2024) proposed PairEval, an LLM-based metric using pairwise comparisons, but it needs fine-tuning on human data and is limited to turn-level evaluation. Other studies (Lin and Chen, 2023; Liu et al., 2023; Chan et al., 2024; Fu et al., 2024) examine unsupervised prompting, mostly with closed models, raising reproducibility and reliability concerns (La Malfa et al., 2024). In contrast, DIET leverages open-source models.

FED (Mehri and Eskénazi, 2020) and FULL (De Bruyn et al., 2022) assess dialogue quality by evaluating how likely a language model is to continue a conversation with a fixed set of follow-ups. FED uses DialoGPT (Zhang et al., 2020) and joint log-likelihood to score 18 dialogue qualities, with positive and negative follow-ups tuned on 10 FED dialogues (Mehri and Eskénazi, 2020). FULL extends FED by computing the conditional log-likelihood of negative follow-ups using the Blender model (Roller et al., 2021), conditioned on dialogue history, selecting them via correlation with FED human scores. While promising, both depend on annotated data for follow-up definition and a fixed set of utterances across all dialogues, limiting adaptability (Kawamoto et al., 2023). DIET overcomes these limitations by generating and selecting dialogue-specific follow-ups without requiring annotations.

3. DIET

This work aims to automate the evaluation of dialogue quality at both the dialogue and turn levels through the use of follow-up utterances. Leveraging LLMs with a meta-prompting strategy, we generate insightful and context-specific follow-ups. Additionally, to improve evaluation, we instruct the models

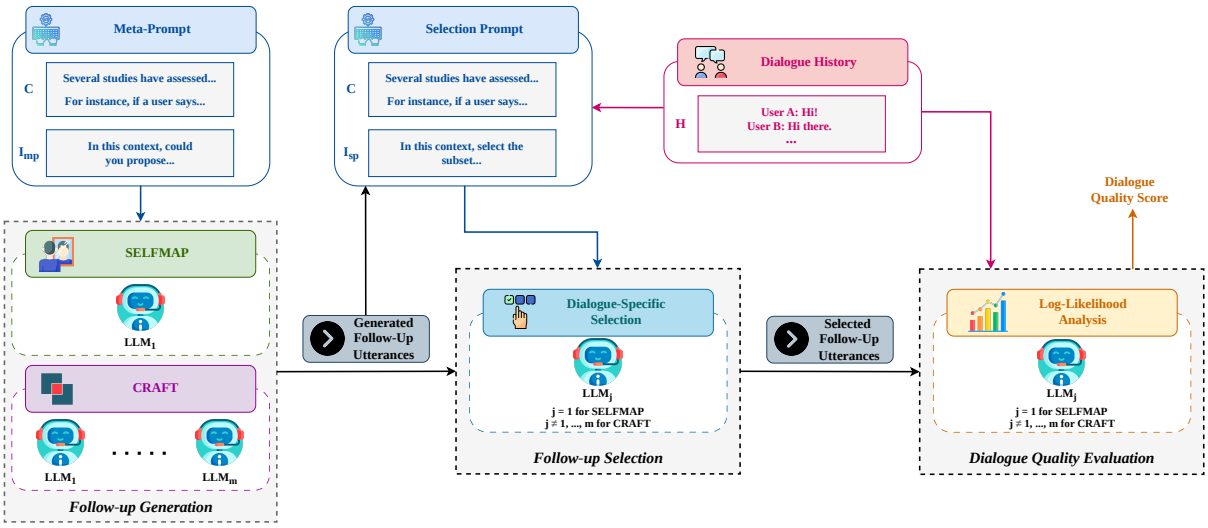

Figure 2: Overview of DIET methodology: (i) *Follow-up generation* via SELFMAP or CRAFT; (ii) *Follow-up selection* based on the dialogue history; and (iii) *Dialogue quality evaluation*.

to identify and select the subset of follow-ups that best assess dialogue quality, rather than utilizing all available follow-ups, as done in FED and FULL. The proposed methodology, referred to as DIET, follows three steps (see Figure 2), detailed in the next subsections.

3.1. Follow-Up Generation Strategies

We propose two meta-prompting strategies: SELFMAP, where one model handles both follow-up generation and evaluation, and CRAFT, which uses multiple models for generation and another for quality assessment. These strategies probe compositionality, defined in Mehri and Eskénazi (2020) as the LLM ability to generate follow-ups even if unobserved during training. We examine this by using familiar follow-ups from the same LLM in SELFMAP and rarer ones from different models in CRAFT.

Meta-Prompt Definition A meta-prompt[2] (Mirza et al., 2024) guides an LLM to generate a set of follow-ups $F = \{f_1, \ldots, f_n\}$, comprising: (i) a context C, defining the function and intended use of follow-ups for dialogue quality evaluation, and (ii) an instruction I_{mp} for follow-up generation. Notably, although human intervention remains necessary for follow-up generation, its extent is substantially minimized. In fact, both C and I_{mp} are defined once and remain invariant across domains, ensuring broad applicability.

SELFMAP This strategy uses follow-ups generated by a language model itself, keeping analysis within its generative scope and reducing reliance on handling novel or rare utterances. To achieve this, the model used for follow-up generation (LLM_1 in Figure 2) is also employed to evaluate dialogue quality within the given history.

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 validate the compositionality property between different 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. Thus, this strategy uses an ensemble of LLMs ($LLM_1, \ldots, LLM_m$) for generating follow-ups, with each LLM receiving the same meta-prompt. Follow-ups from all models are merged, with a separate LLM (LLM_j, where $j \neq 1, \ldots, m$) used for dialogue quality evaluation.

Negative Follow-Ups Building on previous findings that negative follow-ups better align with human evaluations (De Bruyn et al., 2022), we focus on such follow-ups, exemplified by statements like "*Not really relevant here*" and "*What are you trying to say?*". To guide LLMs in generating these follow-ups, we include a negative example into the context C of the meta-prompt.

3.2. Follow-Up Selection Process

Selecting relevant follow-ups is crucial for robust dialogue evaluation. Previous methods like FED and FULL rely on fixed, predefined follow-ups, which, while consistent, lack adaptability to dialogue nu-

[2]The meta-prompt employed for follow-up generation is presented in Appendix A.

ances and may overlook context-specific issues. To address this limitation, we propose a process where the evaluation model selects the most relevant follow-ups, ensuring the selected utterances are more interpretable and contextually appropriate for the given dialogue.

Given a set F of n follow-ups from SELFMAP or CRAFT strategies and a dialogue history H, the goal is to select a subset $F^*_{\mathrm{H}} \subseteq F$ that best reflects the overall quality of H. An evaluator model, LLM_{j}, filters the most informative follow-ups using a selection prompt[3], which shares the context C with the meta-prompt but includes a different instruction I_{sp}. The process adapts to the evaluation level, refining choices based on the scope of analysis. At the turn level, the prompt evaluates the last response, highlighting critical aspects to capture localized inconsistencies. At the dialogue level, the prompt evaluates the full conversation, selecting follow-ups that best represent key conversation traits.

3.3. Dialogue Quality Evaluation via Log-Likelihood

DIET evaluates the quality of a dialogue H using an LLM and d follow-ups generated via SELFMAP or CRAFT, which are filtered through the selection process. For each follow-up f_{i}, DIET computes the log-likelihood that LLM_{j} generates f_{i} from H, denoted as $G_{\mathrm{LLM_j}}(H, f_{\mathrm{i}})$. The overall quality score is the average of these log-likelihoods:

$$QualityScore_{\mathrm{H}} = \frac{1}{d}\sum_{i=1}^{d} G_{\mathrm{LLM_j}}(H, f_{\mathrm{i}})$$

As the evaluation relies on negative follow-ups, a higher log-likelihood $G_{\mathrm{LLM_j}}(H, f_{\mathrm{i}})$ indicates a more probable follow-up f_{i} given dialogue history H, implying lower dialogue quality. We therefore use negative log-likelihood to align scores with human judgments. The formula computes dialogue-level scores but can be adapted to assess a single response within its dialogue history. This is achieved by incorporating the response r into the formula, resulting in $G_{\mathrm{LLM_j}}(H, r, f_{\mathrm{i}})$.

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.

[3]The selection prompt employed for follow-up selection is presented in Appendix A.

4. Experimental Setup

4.1. Considered LLMs

We use open-source models for dialogue scoring, as proprietary models lack probability analysis capabilities. Studies show that LLMs fine-tuned on conversational data more effectively capture dialogue quality nuances (Mehri and Eskénazi, 2020; De Bruyn et al., 2022; Cimino et al., 2024), whereas instruction-tuned models play a key role in follow-up generation and selection. Thus, we employ the following open-source chat models optimized for dialogues: `Chatglm3-6B-Base` (GLM et al., 2024), `Llama-2-13B-Chat` (Touvron et al., 2023), `Qwen-14B-Chat` (Bai et al., 2023), `Vicuna-13B-v1.5` (Chiang et al., 2023), and `Baichuan2-13B-Chat` (Yang et al., 2023). We chose these models to ensure fair comparison across diverse architectures, focusing on model sizes that are well-represented in the open-source landscape. In SELFMAP, a single model handles both follow-up generation and dialogue evaluation. In CRAFT, four models generate follow-ups, and a fifth evaluates, yielding five unique model pairings. Regardless of generation strategy, the evaluation model is also involved in the selection process.

Meta and selection prompts were defined without tuning on annotated data to avoid bias (Kawamoto et al., 2023). The meta-prompt was refined by checking that generated follow-ups met dialogue quality criteria (Mehri and Eskénazi, 2020), while the selection prompt was refined by ensuring chosen follow-ups fit an unlabeled out-of-distribution dialogue set not used in the evaluation process.

To mitigate stochastic variability, selection runs three times per dialogue, randomizing follow-up order to reduce position bias (Shi et al., 2024). The final set is the intersection of chosen subsets; if empty, all follow-ups are used. This was rare, only 45 out of $2,325$ dialogue-level and 10 out of $1,035$ turn-level evaluations had empty intersections.

4.2. Meta-Evaluation Datasets

We evaluate DIET by comparing its scores to human annotations across four meta-evaluation datasets: **FED** (Mehri and Eskénazi, 2020), with human-system and human-human dialogues scored 0–4 at both turn and dialogue levels; **DSTC9** (Gunasekara et al., 2020), featuring conversations between participants and knowledge-grounded generation models, annotated only at the dialogue-level with scores ranging from 1 to 5; **Topical-Chat** (Mehri and Eskenazi, 2020), with multi-topic human-human dialogues scored 0–5 at the turn level; and **Persona-Chat** (Mehri and Eskenazi, 2020), including human-human dialogues where participants are instructed to base their responses on a given per-

sona, also annotated exclusively at the turn-level with scores ranging from 0 to 5.

4.3. Evaluation Metrics

We evaluate the correlation between predicted quality scores and human annotations using Spearman ρ and Pearson r correlation coefficients. Furthermore, the statistical significance of these measures is assessed by computing their respective p-values.

4.4. Compared Methods

We assess the performance of DIET in comparison to 13 unsupervised reference-free metrics.

5. Results

5.1. DIET Performance

Table 1 shows correlation coefficients for all DIET configurations, evaluated across meta-prompting strategies (SELFMAP vs. CRAFT) and log-likelihood methods (joint (JLL) vs. conditional (CLL)) at both dialogue and turn levels.

Overall, CLL consistently outperforms JLL; plausibly because CLL conditioning follow-up probabilities on dialogue history rather than modeling entire sequences, reducing biases from extended contexts and better capturing response relevance. JLL, in contrast, may favor shorter responses due to length bias. Notably, SELFMAP-CLL and CRAFT-CLL show stronger correlations than their JLL variants, reinforcing CLL's effectiveness across meta-prompting strategies.

Table 1 shows CRAFT usually matches or outperforms SELFMAP, except with `Chatglm-3-6B-Base`, where SELFMAP is slightly better. This suggests multiple models enhance follow-up diversity and dialogue assessment, but model size matters: the smallest model, `Chatglm-3-6B-Base`, benefits from judging its own outputs, making compositionality less critical. Larger, more expressive LLMs gain more from CRAFT, implying compositionality becomes increasingly important as models scale.

The selection of the underlying LLM significantly impacts dialogue quality evaluation, as models vary in alignment with human judgments. Despite all selected open-source models being optimized for conversation, differences in training data, architecture, and fine-tuning impact performance. On average, `Vicuna-13B-v1.5` achieves the highest correlation with human scores at both dialogue and turn levels, making it the most suitable for this task.

5.2. Reference-Free Metric Comparison

Table 2 compares DIET's performance with 13 unsupervised reference-free metrics. The upper section presents metrics based on open-source approaches, while the lower section includes closed-source approaches, specifically GPT-Score (Fu et al., 2024) with `GPT-3` and G-Eval (Liu et al., 2023) with `GPT-4`.

DIET consistently outperforms most baselines across datasets and correlation metrics using open-source models. At the dialogue level, only FULL surpasses DIET on the FED dataset, though its follow-ups were selected from FED annotations, likely biasing results. At the turn level, FULL performs comparably to DIET on FED, but underperforms on other datasets. Among other baselines, USL-H (Phy et al., 2020) is competitive at the turn level but weak at the dialogue level. Overall, DIET achieves the highest average correlation across both levels.

For proprietary LLM-based metrics, we were unable to derive correlation scores for GPT-Score, except for the FED dataset (results provided in its respective paper (Fu et al., 2024)), due to the deprecation of GPT-3's davinci01 version. Conversely, G-Eval, based on `GPT-4`, was run on all datasets. Despite DIET's `Vicuna-13B-v1.5` being far smaller than GPT-Score's `GPT-3` (13B vs. 175B), DIET outperforms GPT-Score on FED at both dialogue and turn levels. On the other hand, G-Eval slightly surpasses DIET at the dialogue level, with a more significant advantage at the turn level. However, G-Eval relies on a model of vastly greater scale (13B versus 1.76T parameters) and raises concerns regarding reliability and generalizability due to `GPT-4`'s proprietary nature. Metrics derived from closed-source, API-accessible models inherently introduce uncertainties about their true evaluative capabilities (La Malfa et al., 2024). A key issue is data opacity, as the fine-tuning datasets are not disclosed, creating the potential for data contamination if these models were exposed to the same benchmarks used in our study. Additionally, API-based models exhibit instability, as frequent updates or deprecations by providers compromise reproducibility and long-term validity, as observed with GPT-Score. Thus, DIET remains a lean, transparent alternative for dialogue quality evaluation.

5.3. Qualitative Analysis

DIET accurately identifies responses with clear flaws, like the repeated phrase "*I think he's a great player. I think he's a great player.*" in Table 3, scoring 2.15 close to the human score of 2. This alignment reflects DIET's ability to detect key shortcomings: (i) repetition without new information; (ii) lack of contextual relevance, failing to address prior discussion points (e.g., Jon Gruden's role and Khalil Mack's trade); and (iii) disengagement by neither acknowledging the preceding question nor prompting further interaction.

DIET's LLMs and Methods	Dialogue-level FED ρ	Dialogue-level FED r	Dialogue-level DSTC9 ρ	Dialogue-level DSTC9 r	Dialogue-level Avg. ρ	Dialogue-level Avg. r	Turn-level FED ρ	Turn-level FED r	Turn-level Persona-Chat ρ	Turn-level Persona-Chat r	Turn-level Topical-Chat ρ	Turn-level Topical-Chat r	Turn-level Avg. ρ	Turn-level Avg. r
Chatglm3-6B-Base														
SELFMAP-JLL	0.25	0.20	0.10	0.06	0.18	0.13	0.09*	0.08*	−0.10*	−0.12	−0.10*	−0.10	−0.04	−0.05
CRAFT-JLL	0.25	0.20	0.10	0.06	0.18	0.13	0.09*	0.08*	−0.10*	−0.12	−0.09*	−0.10*	−0.03	−0.05
SELFMAP-CLL	0.57	0.52	0.14	0.13	0.36	0.33	0.45	0.42	0.05*	0.06*	0.09*	0.08*	0.20	0.19
CRAFT-CLL	0.53	0.48	0.09	0.08	0.31	0.28	0.39	0.32	0.02*	0.05*	0.22	0.23	0.21	0.20
Llama-2-13B-Chat														
SELFMAP-JLL	0.25	0.21	0.11	0.07	0.18	0.14	0.10	0.09*	−0.10*	−0.11	−0.10*	−0.11	−0.03	−0.04
CRAFT-JLL	0.25	0.20	0.11	0.07	0.18	0.14	0.10	0.09*	−0.10*	−0.11*	−0.10*	−0.11	−0.03	−0.04
SELFMAP-CLL	0.54	0.46	0.10	0.08	0.32	0.27	0.52	0.47	0.07*	0.09*	0.04*	0.04*	0.21	0.20
CRAFT-CLL	0.54	0.53	0.13	0.13	0.34	0.33	**0.54**	**0.52**	0.07*	0.08*	0.09*	0.08*	0.23	0.23
Qwen-14B-Chat														
SELFMAP-JLL	0.20	0.19	0.09	0.06	0.15	0.13	0.08*	0.08*	−0.08*	−0.09*	−0.09*	−0.09*	−0.03	−0.03
CRAFT-JLL	0.20	0.19	0.09	0.06	0.15	0.13	0.07*	0.07*	−0.08*	−0.09*	−0.08*	−0.09*	−0.03	−0.04
SELFMAP-CLL	0.46	0.39	0.15	0.13	0.31	0.26	0.39	0.32	0.15	0.11*	0.20	0.16	0.25	0.20
CRAFT-CLL	0.51	0.45	0.16	0.15	0.34	0.30	0.45	0.40	0.23	0.24	0.20	0.19	0.29	0.28
Vicuna-13B-v1.5														
SELFMAP-JLL	0.27	0.21	0.11	0.07	0.19	0.14	0.09*	0.09*	−0.10*	−0.11*	−0.10*	−0.11	−0.04	−0.04
CRAFT-JLL	0.28	0.22	0.11	0.07	0.20	0.15	0.10*	0.09*	−0.09*	−0.10*	−0.10*	−0.10	−0.03	−0.04
SELFMAP-CLL	0.52	0.46	0.12	0.11	0.32	0.29	0.45	0.41	0.14	0.13	0.15	0.13	0.25	0.22
CRAFT-CLL	**0.63**	**0.58**	**0.19**	**0.18**	**0.41**	**0.38**	0.52	0.46	**0.28**	**0.27**	0.29	0.28	**0.36**	**0.34**
Baichuan2-13B-Chat														
SELFMAP-JLL	0.27	0.21	0.12	0.04*	0.20	0.13	0.11	0.09*	−0.09*	−0.11*	−0.09*	−0.10*	−0.02	−0.04
CRAFT-JLL	0.26	0.20	0.11	0.04*	0.19	0.12	0.09*	0.08*	−0.10*	−0.11*	−0.09*	−0.10*	−0.03	−0.04
SELFMAP-CLL	0.51	0.42	0.18	0.15	0.35	0.29	0.34	0.32	0.09*	0.10*	0.19	0.15	0.21	0.19
CRAFT-CLL	0.57	0.50	0.17	0.16	0.37	0.33	0.47	0.47	0.16	0.15	**0.30**	**0.30**	0.31	0.31

Table 1: Correlation coefficients between human evaluations and DIET scores, using SELFMAP or CRAFT for follow-up generation and the follow-up selection process. ρ and r represent Spearman's rank correlation coefficient and Pearson correlation, respectively. Bold values indicate the highest scores, while underlined values mark the best-performing configuration among the four settings for each model. All values with $p > 0.05$ are marked with an asterisk (*). JLL: Joint Log-Likelihood. CLL: Conditional Log-Likelihood.

Metric	Dialogue-level FED ρ	Dialogue-level FED r	Dialogue-level DSTC9 ρ	Dialogue-level DSTC9 r	Dialogue-level Avg. ρ	Dialogue-level Avg. r	Turn-level FED ρ	Turn-level FED r	Turn-level Persona-Chat ρ	Turn-level Persona-Chat r	Turn-level Topical-Chat ρ	Turn-level Topical-Chat r	Turn-level Avg. ρ	Turn-level Avg. r
Metrics based on open-source approaches														
DIET (Ours)	0.63	0.58	**0.19**	**0.18**	**0.41**	**0.38**	**0.52**	0.46	0.28	0.27	0.29	0.28	**0.36**	**0.34**
FED (Mehri and Eskénazi, 2020)	0.32	0.22	0.12	0.13	0.22	0.18	0.10	0.12	0.00*	−0.03*	−0.14	−0.12	−0.01	−0.01
FULL (De Bruyn et al., 2022)	**0.69**	**0.64**	0.10	0.12	0.40	**0.38**	0.51	**0.47**	0.09	0.07	−0.07*	−0.05*	0.18	0.16
USL-H (Phy et al., 2020)	0.15*	0.07*	0.11	0.11	0.13	0.09	0.19	0.20	**0.52**	**0.50**	**0.34**	0.32	0.35	**0.34**
USR (Mehri and Eskenazi, 2020)	0.06*	0.09*	0.02*	0.02*	0.04	0.06	0.12	0.11	0.42	0.44	0.33	**0.41**	0.29	0.32
FlowScore (Li et al., 2021)	0.00*	−0.07*	0.14	0.15	0.07	0.04	−0.06*	−0.07*	0.08*	0.12*	0.08*	0.10*	0.03	0.05
DynaEval (Zhang et al., 2021)	0.55	0.50	0.10	0.09	0.33	0.30	0.32	0.32	0.17	0.15	−0.02*	−0.03*	0.16	0.15
DEB (Sai et al., 2020)	0.00*	−0.13*	0.13	0.09	0.07	−0.02	0.19	0.23	0.37	0.29	0.12	0.18	0.23	0.23
GRADE (Huang et al., 2020b)	−0.07*	−0.03*	−0.07	−0.08	−0.07	0.03	0.12	0.13	0.35	0.36	0.22	0.20	0.23	0.23
DEnsity (Park et al., 2023)	–	–	–	–	–	–	0.21	0.25	0.35	0.36	0.25	0.16	0.27	0.26
RoBERTa-eval (Zhao et al., 2020)	–	–	–	–	–	–	0.26	0.29	0.33	0.34	0.22	0.22	0.27	0.28
QualityAdapt (Mendonça et al., 2022)	–	–	–	–	–	–	0.23	0.27	0.36	0.36	0.29	0.29	0.29	0.31
Metrics based on closed-source approaches														
GPT-Score (Fu et al., 2024)	0.54	–	–	–	–	–	0.38	–	–	–	–	–	–	–
G-Eval (Liu et al., 2023)	0.61	0.64	0.22	0.24	0.42	0.44	0.48	0.56	0.58	0.57	0.56	0.64	0.54	0.59

Table 2: Comparison of DIET scores with unsupervised reference-free metrics. Bold values indicate the highest scores among metrics that use open-source approaches. All values with $p > 0.05$ are marked with an asterisk (*).

We observe some challenges in evaluating dialogues involving pragmatic subtleties. For example, in the conversation shown in Table 4, the system produces witty responses like "*I see you are one of those people who writes 'Hmm' with 2 m's*" and "*1 2 3 4 5 6 7 8 9 ten*", maintaining conversational flow. While human annotators rated this exchange relatively high (2.4), DIET assigned a lower score (0.39), reflecting its stronger focus on aspects such as literal clarity and topical relevance. This shows DIET's limits with extremely non-literal language, where tone and playfulness prevail. However, this limitation will be mitigated as future LLMs improve their ability to interpret pragmatic subtleties like humor (Jentzsch and Kersting, 2023; Hu et al., 2024).

5.4. Meta-Correlation Analysis

To assess whether standard language model benchmarks predict LLM effectiveness in unsupervised dialogue evaluation, we perform a meta correlation analysis. We examine whether LLM performance on general purpose benchmarks, MMLU (Hendrycks et al., 2020), AGIEval (Zhong et al., 2024), and BBH (Suzgun et al., 2023), correlates with alignment between DIET scores and human

Dialogue History
Speaker A: I think that Jon Gruden is a sensible loss for the team.
Speaker B: What do you mean by sensible loss? I think Jon Gruden is a huge mistake for the team.
Speaker A: Well, I think he's a good player, hard to replace, that's what I meant - or at least in the Raiders team he's become an important piece for the coach.
Speaker B: Jon Gruden is an NFL head coach. He's not a player. The 33-13 loss to the Rams was just the beginning of a miserable season. Are you familiar with the NFL?
Speaker A: Oh, I see. I guess I'm confusing him with a player on the team. What was his name? I think it was Khalil Mack. And I do follow some games—what about you?
Speaker B: Football is one of my favorite sports, and I follow all 32 teams. Khalil Mack was traded to the Bears by Jon Gruden for some picks. They definitely felt his absence. I don't know why they didn't just re-sign him.
Speaker A: Maybe he demanded a high salary? What is your favorite NFL team?
Response: I think he's a great player. I think he's a great player.
Human Score: Overall Score [3, 2, 1], Average: *2*
DIET Score: *2.15*

Table 3: A dialogue history and its corresponding response from Topical-Chat, along with the human score (on a 0–5 scale) and DIET's min-max scaled prediction.

Dialogue History
User: Hi!
System: Hi there.
User: Can you help me?
System: What kind of help would you like?
User: Hmm, like booking a ticket or ordering food?
System: I see you are one of those people who writes "Hmm" with 2 m's. Interesting comparison.
User: Oh, you know, to count?
System: 1 2 3 4 5 6 7 8 9 ten.
User: Lol, what's after that?
System: I am glad you find this funny. The successor to that.
User: Okay, tell me what comes after 10?
Human Score: Overall Score [3, 3, 2, 2, 2], Average: *2.4*
DIET Score: *0.39*

Table 4: A dialogue history from FED with its human score (on a 0–4 scale) and DIET's predicted score, which was scaled to the same range using min-max normalization.

ID	Follow-Up
1	I don't think that's a very good response.
2	That's not the answer I was hoping for.
3	That's not very helpful/relevant at all.
4	That's not what I meant.
5	I don't think that's the right answer.
6	I'm not sure I follow your reasoning.
7	Oh, really? Why is that interesting?
8	Sorry, I didn't quite catch that. Could you please clarify?
9	That's an interesting point, but how does it relate to our conversation so far?
10	I see, but how is that related to my question/comment?
11	Your answer is not quite what I had in mind when I made my comment.
12	That's an interesting fact, but it doesn't answer my query.

Table 5: Follow-up utterances employed by `Vicuna-13B-v1.5` with CRAFT-CLL strategy.

judgments. DIET scores are computed as the mean of Spearman's ρ and Pearson's r across dialogue and turn level evaluations, using averages from the DIET's optimal configuration for each model.

Table 6 reports open-source chat model rankings across benchmarks and their DIET-based alignment with human evaluations. Results show no consistent relationship between benchmark scores and human-alignment: models scoring lower on MMLU, AGIEval, and BBH, like `Vicuna-13B-v1.5`, can achieve the highest alignment with human judgments, while top-scoring models, such as `Chatglm3-6B-Base`, perform poorly. Furthermore, models that perform well on one benchmark exhibit weak performance on others, underscoring the inconsistency in the predictive value of these benchmarks. Task-specific evaluations are therefore essential when selecting LLMs for dialogue quality assessment.

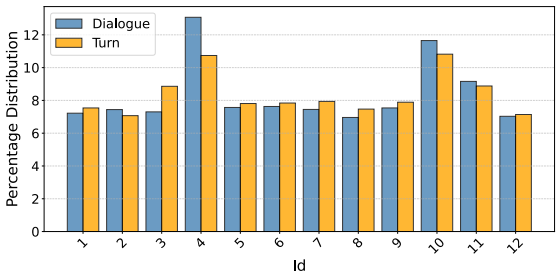

Figure 3: Percentage distribution of follow-up usage by `Vicuna-13B-v1.5` with CRAFT-CLL at both dialogue and turn levels.

Model	DIET (%)	MMLU (%)	AGIEval (%)	BBH (%)
Vicuna-13B-v1.5	**1** (37)	5 (55)	5 (32)	5 (43)
Baichuan2-13B-Chat	2 (33)	3 (59)	3 (48)	3 (49)
Qwen-14B-Chat	3 (30)	**1** (65)	2 (52)	2 (53)
Llama-2-13B-Chat	4 (28)	4 (55)	4 (32)	4 (47)
Chatglm3-6B-Base	5 (25)	2 (61)	**1** (54)	**1** (66)

Table 6: Comparison of DIET performance rankings and benchmark scores for each LLM (higher is better). The leading index in each cell denotes the model's rank for the corresponding metric, while the value in parentheses indicates its actual score.

6. Further Investigation

6.1. Follow-Up Analysis

This section analyzes the role of follow-ups in dialogue quality evaluation, focusing on the most frequently selected ones at dialogue and turn levels under DIET's optimal configuration. The full set of follow-ups is listed in Table 5.

As shown in Figure 3, the most frequently chosen follow-ups for dialogue-level evaluation indicate misinterpretation or topic irrelevance across multiple turns. The high selection of Follow-Up 4 suggests that dialogues with consistent misinterpretations of user intent receive lower quality scores. Follow-Up 10 emphasizes concerns about topic coherence, which is critical for evaluating conversational continuity and logical flow. Follow-Up 11 captures cases where responses fail to meet user expectations across exchanges, reinforcing DIET's

#Follow-ups	Dialogue-level						Turn-level							
	FED		DSTC9		Avg.		FED		Persona-Chat		Topical-Chat		Avg.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
12	**0.63**	**0.58**	**0.19**	**0.18**	**0.41**	**0.38**	**0.52**	**0.46**	**0.28**	**0.27**	**0.29**	**0.28**	**0.36**	**0.34**
8	0.59	0.57	0.14	0.16	0.37	0.37	0.50	0.45	0.26	0.26	0.26	0.25	0.34	0.32
4	0.56	0.54	0.11	0.12	0.34	0.33	0.48	0.45	0.17	0.19	0.22	0.20	0.29	0.28

Table 7: Effect of the number of follow-ups on evaluation performance using the Vicuna-CRAFT-CLL configuration. ρ and r represent Spearman's rank correlation coefficient and Pearson correlation, respectively. Higher correlation values indicate stronger alignment with human judgments. Bold values indicate the highest scores.

DIET's LLMs and Methods	Dialogue-level						Turn-level							
	FED		DSTC9		Avg.		FED		Persona-Chat		Topical-Chat		Avg.	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
Chatglm3-6B-Base														
SELFMAP-JLL	0.23	0.18	0.10	0.06	0.17	0.12	0.07*	0.07*	−0.11*	−0.12	−0.10*	−0.11	−0.05	−0.05
CRAFT-JLL	0.22	0.18	0.10	0.06	0.16	0.12	0.07*	0.07*	−0.11*	−0.12	−0.10*	−0.10	−0.05	−0.05
SELFMAP-CLL	0.56	0.54	0.06	0.05	0.31	0.30	0.44	0.47	−0.13	−0.11	−0.07*	−0.07*	0.08	0.10
CRAFT-CLL	0.47	0.43	−0.01*	−0.01*	0.23	0.21	0.31	0.36	−0.08*	−0.05*	0.18	0.19	0.14	0.17
Llama-2-13B-Chat														
SELFMAP-JLL	0.24	0.19	0.11	0.07	0.18	0.13	0.08*	0.08*	−0.10*	−0.12	−0.10*	−0.11	−0.04	−0.05
CRAFT-JLL	0.24	0.19	0.10	0.07	0.17	0.13	0.08*	0.07*	−0.10*	−0.12	−0.10*	−0.10	−0.04	−0.05
SELFMAP-CLL	0.52	0.52	0.05	0.07	0.29	0.30	0.44	0.48	0.06*	0.07*	−0.03*	−0.04*	0.16	0.17
CRAFT-CLL	**0.59**	**0.59**	0.05	0.07	**0.32**	**0.33**	**0.49**	**0.53**	0.05*	0.08*	−0.04*	−0.05*	0.17	0.19
Qwen-14B-Chat														
SELFMAP-JLL	0.18	0.18	0.09	0.06	0.14	0.12	0.05*	0.06*	−0.09*	−0.10*	−0.09*	−0.10*	−0.04	−0.05
CRAFT-JLL	0.19	0.18	0.09	0.06	0.14	0.12	0.05*	0.06*	−0.09*	−0.10*	−0.09*	−0.10*	−0.04	−0.05
SELFMAP-CLL	0.48	0.44	0.11	**0.12**	0.30	0.28	0.34	0.38	**0.16**	**0.15**	0.22	0.21	0.24	0.25
CRAFT-CLL	0.53	0.49	0.08	0.09	0.31	0.29	0.40	0.42	0.15	**0.15**	0.18	0.17	0.24	0.25
Vicuna-13B-v1.5														
SELFMAP-JLL	0.25	0.20	0.11	0.07	0.18	0.14	0.08*	0.08*	−0.10*	−0.12	−0.10*	−0.11	−0.04	−0.05
CRAFT-JLL	0.25	0.20	0.11	0.07	0.18	0.14	0.08*	0.08*	−0.10*	−0.11	−0.10*	−0.11	−0.04	−0.05
SELFMAP-CLL	0.47	0.46	0.07	0.08	0.27	0.27	0.44	0.47	0.06*	0.05*	0.11	0.10	0.20	0.21
CRAFT-CLL	0.55	0.53	0.08	0.09	**0.32**	0.31	0.47	0.50	0.12	**0.15**	0.14	0.14	0.24	0.26
Baichuan2-13B-Chat														
SELFMAP-JLL	0.24	0.19	0.11	0.04*	0.18	0.12	0.08*	0.07*	−0.10*	−0.12	−0.09*	−0.10	−0.04	−0.05
CRAFT-JLL	0.24	0.19	0.11	0.04*	0.18	0.12	0.08*	0.07*	−0.10*	−0.12	−0.09*	−0.10	−0.04	−0.05
SELFMAP-CLL	0.50	0.51	**0.14**	0.04*	**0.32**	0.28	0.38	0.43	0.06*	0.07*	0.16	0.15	0.20	0.22
CRAFT-CLL	0.52	0.51	0.08	0.05	0.30	0.28	0.44	0.48	0.08*	0.10*	**0.24**	**0.24**	**0.25**	**0.27**

Table 8: Correlation coefficients between human evaluations and DIET scores, using SELFMAP or CRAFT for follow-up generation without the application of the follow-up selection process. Bold values indicate the highest scores, while underlined values mark the best-performing configuration among the four settings for each model. All values with $p > 0.05$ are marked with an asterisk (*).

ability to detect long-range coherence issues.

At the turn level, evaluation focuses on immediate response quality. In Figure 3, frequent selection of Follow-Up 10 shows that response relevance is a key factor in turn-level assessments. Follow-Up 4 reveals how misinterpretations hurt quality, while Follow-Up 3 shows penalties for responses lacking useful or contextually appropriate information are heavily penalized. Similarly, Follow-Up 11 highlights the importance of aligning system responses with user intent, as misalignments notably degrade perceived quality.

6.2. Impact of Follow-Up Pool Size

To assess how the number of candidate follow-ups affects evaluation, we ran supplementary experiments using our best setup, `Vicuna-CRAFT-CLL`. We varied the initial follow-up pool before selection: 12 (three per model), 8 (two per model), and 4 (one per model), sampling utterances randomly from each model. Table 7 shows that more candidate follow-ups improve performance at both dialogue and turn levels. These findings suggest that a larger, more diverse pool strengthens selection, yielding more accurate assessments. While results may improve with even more follow-ups, this also raises the computational cost of log-likelihood evaluation for each additional candidate.

6.3. Impact of Follow-Up Selection Process

We further conduct an ablation study on the impact of selection process in DIET's performance. As shown in Table 8, the results reveal that performance without selection is significantly lower compared to the outcomes achieved using dialogue-specific follow-ups (see Table 1). Without selection, CLL includes irrelevant follow-ups, reducing alignment with human scores, while JLL remains more stable using full dialogue history. This highlights the importance of filtering for CLL. Notably, without selection, `Llama-13B-Chat` and `Baichuan2-13B-Chat` lead at dialogue and turn levels, but with selection, `Vicuna-13B-v1.5` performs best.

7. Conclusion

This paper introduces DIET, a novel unsupervised and reference-free metric for dialogue quality evaluation. Among the two meta-prompting strategies, we found that CRAFT consistently performs better than SELFMAP, highlighting the effectiveness of multiple LLMs in generating diverse and complementary follow-ups. Comparing the two log-likelihood computation methods, the results show that CLL outperforms JLL across all datasets. Crucially, the follow-up selection process proves essential to maximizing alignment with human judgments, especially under the CLL setting. Encouragingly, facing much larger proprietary models like GPT-3 and 4, our DIET outperforms GPT-Score on FED at both dialogue and turn levels.

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 $LLM_1, \ldots, LLM_{\mathrm{m}}$ 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.

A current limitation of DIET is its capacity to compute only a single overall quality score. While this provides useful insights into the general quality of a dialogue, it lacks the granularity required for evaluating specific attributes of a conversation. Future work will aim to enhance DIET by incorporating the evaluation of multiple quality dimensions, such as coherence, relevance, and engagement. This extension will facilitate a more comprehensive assessment, allowing practitioners to identify areas for improvement in domain-specific dialogues.

Finally, we assessed DIET in open-domain conversations. To adapt DIET for domain-oriented applications, a promising approach involves enabling the generation of domain-specific follow-ups. By integrating domain-specific knowledge, DIET could be modified to produce follow-ups that are not only contextually relevant but also aligned with the nuances of particular domains. This extension could enhance the model's ability to assess dialogue quality in specialized contexts, such as customer support or healthcare, where domain knowledge plays a crucial role in the interaction.

9. Ethics Statement

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 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.

10. Bibliographical References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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.

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 *Proceedings of the Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

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 and Vincenzo Deufemia. 2025. Sigfrid: Unsupervised, platform-agnostic interference detection in iot automation rules. *ACM Transactions on Internet of Things*, 6(2):1–33.

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, SIGDIAL 2024, Kyoto, Japan, September 18-20, 2024*, pages 297–316. Association for Computational Linguistics.

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.

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

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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. *arXiv preprint arXiv:2011.06486*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024. Cracking the code of juxtaposition: Can AI models understand the humorous contradictions. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020a. 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.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020b. 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)*, pages 9230–9240, Online. Association for Computational Linguistics.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

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.

Emanuele La Malfa, Aleksandar Petrov, Simon Frieder, Christoph Weinhuber, Ryan Burnell, Raza Nazar, Anthony Cohn, Nigel Shadbolt, and Michael Wooldridge. 2024. Language-models-as-a-service: Overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research*, 80:1497–1523.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv: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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. 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. *arXiv preprint arXiv: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.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. 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.

Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Sivan Doveh, Jakub Micorek, Mateusz Kozinski, Hilde Kuehne, and Horst Possegger. 2024. Meta-prompting for automating zero-shot visual recognition with llms. *arXiv preprint arXiv:2403.11755*.

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, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

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.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online). International Committee on 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.

Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

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.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

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.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yi-Ting Yeh, Maxine Eskénazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.

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.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11657–11666. AAAI Press.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A. Meta and Selection Prompt

The meta-prompt, as presented in Figure 4, directs LLMs in generating follow-up utterances for dialogue quality assessment. It comprises a context that clarifies the function of follow-ups in dialogue evaluation and includes an illustrative example, along with an instruction that explicitly guides the model in producing follow-ups. Conversely, the selection prompt, shown in Figure 5, enhances the evaluation process by filtering the generated follow-ups to retain only the most pertinent ones.

Meta-Prompt

Context:
Several studies have assessed the quality of utterances in interactive settings by examining the subsequent user responses. For instance, if a user says, "How are you today? I was just watching the football game", and receives a reply such as, "Did you know cats sleep for 18 hours a day?", a typical follow-up utterance from the user, highlighting the irrelevance of the response, might be, "Huh? That's not relevant at all...".
Instruction:
In this context, could you propose a set of potential follow-up utterances that could be employed to evaluate the quality of responses in general conversations?

Figure 4: Meta-prompt designed to generate follow-ups in dialogue evaluation using the DIET metric.

Selection Prompt

Context:
Several studies have assessed the quality of . . .
Instruction:
In this context, select the subset or the entire set of follow-up responses from the options below that would be most suitable for evaluating the quality of *the last turn in* the following dialogue:

`[Insert dialogue history here]`

Options:

`[Insert follow-up options here]`

Please provide your selection based on the dialogue history.

Figure 5: Selection prompt designed to select follow-ups in turn- and dialogue-level evaluations using the DIET metric. Underlined text is utilized exclusively for turn-level evaluation.