Decision-Making with Deliberation: Meta-reviewing as a Document-grounded Dialogue

Anonymous ACL submission

Abstract

Meta-reviewing is a pivotal stage in the peer-review process, serving as the final step in determining whether a paper is recommended for acceptance. Prior research on meta-reviewing has treated this as a summarization problem over review reports. However, complementary to this perspective, meta-reviewing is a decision-making process that requires weighing reviewer arguments and placing them within a broader context. Prior research has demonstrated that decision-makers can be effectively assisted in such scenarios via dialogue agents. In line with this framing, we explore the practical challenges for realizing dialog agents that can effectively assist meta-reviewers. Concretely, we first address the issue of data scarcity for training dialogue agents by generating synthetic data using Large Language Models (LLMs) based on a self-refinement strategy to improve the relevance of these dialogues to expert domains. Our experiments demonstrate that this method produces higher-quality synthetic data and can serve as a valuable resource towards training meta-reviewing assistants. Subsequently, we utilize this data to train dialogue agents tailored for meta-reviewing and find that these agents outperform *off-the-shelf* LLM-based assistants for this task. Finally, we apply our agents in real-world meta-reviewing scenarios and confirm their effectiveness in enhancing the efficiency of meta-reviewing.

1 Introduction

Peer review is the cornerstone of academic quality control across all scientific disciplines (Ware and Mabe, 2009). In a typical peer review process, a group of experts (the *reviewers*) assesseses the paper at hand, summarizes their opinion in written reports (the *reviews*), and then a meta-reviewer weighs the presented arguments to decide to *accept* or *reject* the paper for publication resulting in a *meta-review* report. This task requires significant effort and expertise by the meta-reviewer, as they need to match arguments, weigh different opinions, and contextualize the reviews within the broader scientific field. The growing load of research output (Chen et al., 2022; Krenn et al., 2023) puts an additional strain on meta-reviewers requiring to arrive at reliable decisions *quickly*. While the assistance of reviewers through means of natural language processing (NLP) has received considerable attention (e.g., Dycke et al., 2023; Kuznetsov et al., 2024), so far, the support for the decision-making process of meta-reviewers is under-investigated despite the apparent practical needs.

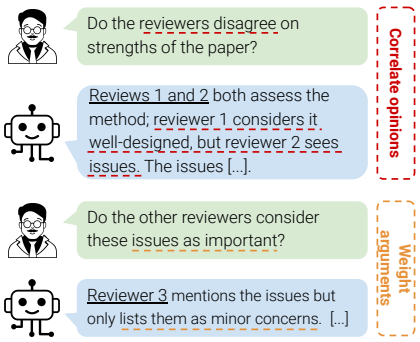


Figure 1: Illustration of the process of *meta-reviewing as a dialogue*. Dialogues include requests to summarize opinions, weight arguments, and contextualize them.

Meta-reviewing has traditionally been framed as a multi-document summarization task (Shen et al., 2022; Li et al., 2023a; Zeng et al., 2023), aimed at generating meta-review reports by condensing the underlying reviews. While these summarization methods can aid meta-reviewers by simplifying the writing process, they often fall short of directly supporting decision-making. A more application-driven and human-centered approach to decision-making remains an important gap.

This paper advocates a new angle on assisting meta-reviewers that focuses on supporting the underlying *decision-making process* to address these needs. For human decision-making and information assimilation over source materials, dialogue assistance has proven effective in many domains

(Lever and Sénéchal, 2011; Golinkoff et al., 2019). Following up on a recent proposal by Balke et al. (2023), we explore the practical implications of developing a dialogue agent that assists the meta-reviewer during their step-wise decision-making process by collaboratively reflecting on the underlying peer reviews. Figure 1 illustrates an example of such a dialogue. Ultimately, we investigate the following research questions. (RQ1) **What requirements does the meta-reviewing domain pose on a dialogue agent?** (RQ2) **How can we develop such a dialogue agent to account for these requirements?** (RQ3) **Is a dialogue agent practically useful during meta-reviewing?**

Developing a meta-reviewing decision-making dialogue assistant presents unique challenges. Unlike prior work in other domains on general topics (Wang et al., 2019; Farag et al., 2022), a meta-reviewing dialogue agent must engage with highly technical topics and ensure its responses are strictly grounded in the underlying reviews. As shown in our experiments in Section 3, off-the-shelf Large Language Models (LLMs) are inadequate for this task, as they fail to meet both these requirements. To bridge this gap, dedicated fine-tuning is crucial to tailor the models for technical grounding and decision-making tasks. Additionally, nuanced evaluation methods are required to ensure their reliability. However, fine-tuning and evaluating such agents is costly, as it demands significant effort to generate high-quality human dialogue data.

To address these issues we develop a dedicated fine-tuning and evaluation suite for meta-reviewing dialogue agents. First, we tackle the lack of training data by leveraging LLMs for **synthetic data generation**, which has shown promise across various fields (e.g., Møller et al., 2023; Chen et al., 2023a). Specifically, we simulate meta-review decision-making dialogues with LLMs such as ChatGPT[1], LLaMa (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Mixtral (Jiang et al., 2024). Through human and automated evaluations, we find that out-of-the-box LLMs generate dialogues that are unspecific and lack diversity. Therefore, we devise a new self-refinement strategy that addresses these challenges in the meta-reviewing domain but also show its generality for other decision-making domains. Afterwards, in Section 4 we utilize the synthetic dataset to **fine-tune dialogue agents** for meta-reviewing assistance and propose an evaluation strategy. Both automatic and human evaluations show that smaller, fine-tuned open-source models outperform larger, closed zero-shot LLMs like ChatGPT in generating relevant and high-quality responses. Finally, in Section 5, we **deploy the best-performing dialogue agent** within a human user study demonstrating the effectiveness of such a dialogue agent in time-efficient meta-reviewing. Our results show that dialogue agents can effectively **reduce the meta-reviewing time by almost up to 50%**, while enabling the creation of more comprehensive and detailed meta-review reports, compared to when no agent is used. This demonstrates the potential of dialogue agents to significantly enhance both the quality and efficiency of the meta-reviewing process.

In summary, our contributions are three-fold: (**1**) We present the first study on meta-reviewing as a dialogue to-date and propose extensive measures towards developing dialogue agents for this scenario (RQ1). (**2**) We address the data scarcity issue for training dialogue agents via synthetic data generation using LLMs. We further propose a self-refinement strategy to enhance the generated dialogues using dialogue quality metrics as feedback (RQ2). (**3**) We utilize the synthetically generated data for training dialogue agents tailored to the meta-reviewing scenario and evaluate the utility of these agents in real-world meta-reviewing (RQ3).

2 Meta-reviewing as a Dialogue

Existing NLP literature treats meta-reviewing exclusively as a summarization problem (Shen et al., 2022; Li et al., 2023a; Zeng et al., 2023). While meta-reviewing involves summarization steps, it is primarily a decision-making process where the human meta-reviewer has to arrive at an overall recommendation on the paper based on the review reports. This decision-making process requires complex reasoning, including weighing, contextualizing, and meta-reasoning on the reviews. For instance, meta-reviewers need to adjust for reviewer uncertainty, biases, and review quality or need to consider the related work and state of the field. Summaries can assist in organizing information but may offer limited direct support for decision-making in meta-reviewing. This is because humans can be influenced by readily presented recommendations, a phenomenon known as *automation bias* (Goddard et al., 2014; Schemmer et al., 2022). In combination with the fact that LLMs

[1] https://chat.openai.com/

sporadically introduce biases and errors in the summaries, this may impact human agency over the process (Donker, 2023; Schintler et al., 2023). Moreover, current ACL policies discourage the use of LLMs as ghostwriters during meta-reviewing.[2]

In response to these issues, we propose a new *human-centered* perspective on meta-reviewing assistance by framing meta-reviewing as a decision-making process that benefits from assistance through a dialogue agent. This concept has shown promise in related domains (Wang et al., 2019; Farag et al., 2022) and in a piloting study by Balke et al. (2023), who simulate this scenario between a human meta-reviewer and a human dialogue assistant. More formally, we define meta-reviewing as a document-grounded decision-making dialogue where a dialogue agent provides information on the underlying reviews, and the meta-reviewer gains insights to move closer to their decision with each turn of the dialogue. While the agent may occasionally introduce biases or errors, the meta-reviewer can verify the information and enhance their understanding through iterative refinement (Ke et al., 2024).[3] Specifically, following Balke et al. (2023), the dialogue agent should *ground* their answers in the reviews without hallucinating or obfuscating information, and their answers should be highly *technically specific*, avoiding generic and unrelated statements. We refer to these two desiderata as GROUNDEDNESS and SPECIFICITY.

Our experiments in Section §3 demonstrate the difficulty of this task. Out-of-the-box LLMs generate answers that fail on both dimensions (GROUNDEDNESS and SPECIFICITY). Thus, dedicated fine-tuning and nuanced evaluation of these agents are essential. However, data for training and evaluating such agents is costly as human experts, and considerable effort is required to generate human-to-human example dialogues as proposed in Balke et al. (2023). We explore these challenges in the remainder of this work.

3 Synthetic Data Generation for Fine-tuning Meta-review Dialogue Agents

Developing effective meta-reviewing dialogue agents necessitates dedicated fine-tuning, as off-the-shelf LLMs fall short of this specialized task (cf. Sec §4). This fine-tuning process relies on high-quality training data tailored to meta-reviewing assistance. However, obtaining expert annotators and producing human gold data is resource-intensive since crowd-sourcing is not a viable option (Stasaski et al., 2020). To address these limitations, leveraging LLMs to generate synthetic data has become a widely adopted approach in expert-driven domains such as education (Macina et al., 2023) and biomedicine (Smolyak et al., 2024). Following this trend, we employ LLMs to *synthetically generate* a dataset of meta-reviewing dialogues, building on their proven utility in various tasks (Møller et al., 2023; Chen et al., 2023a). However, meta-reviewing dialogues are particularly challenging to synthesize, as they need to account for the previously mentioned desiderata of GROUNDEDNESS and SPECIFICITY. Our initial experiments show that out-of-the-box LLMs fail to generate dialogues that meet these criteria. To overcome this, we implement a self-refinement-based approach for data generation. In the following, we present our data generation approach (§3.1), assess dataset quality (§3.3), fine-tune models (§4), and conduct a within-subject evaluation (§5).

3.1 Approach

We propose ReMuSE (**Re**ward-based **Mu**lti-aspect **S**elf-**E**diting) based on the self-refinement paradigm (Madaan et al., 2023; Chen et al., 2023b). In this paradigm, an LLM $\mathcal{M}$ first produces the initial output o. Then, feedback f is obtained on the output via $\mathcal{M}$ or other feedback mechanisms. The model $\mathcal{M}$ is then prompted to correct the initial output o based on f. All prior approaches (Madaan et al., 2023; Chen et al., 2023b) either improve a single aspect of the output only, or use ground truth in-context exemplars to steer the generations towards a particular format. These factors restrict the diversity of the corrections to a small set of specific changes rather than profound revisions (Wang et al., 2023). However, as our experiments show (cf. Section §3.3), meta-reviewing dialogue generation gains from incorporating multiple aspects during the feedback loop of self-refinement. In this work, we thus propose a zero-shot edit strategy that makes use of scoring functions to calculate dialogue quality metrics. These *rewards* are used as feedback to improve the initial output across multiple dimensions.

ReMuSE generates dialogues based on a knowledge source, e.g. the collection of reviews. Given

[2] https://2023.aclweb.org/blog/review-acl23/

[3] We summarize the differences to prior approaches in Table 5

a knowledge source $\mathcal{K}$ and an LLM $\mathcal{M}$ ReMuSE operates in four steps (cf. Figure 3 in §A.2):

I. Initial Dialogue Generation We prompt $\mathcal{M}$ with knowledge source $\mathcal{K}$ to generate an initial dialogue $\mathcal{D}$ in a zero-shot fashion.

II. Evaluation The generated dialogue is then evaluated to obtain rewards using an **evaluator**. The evaluator consists of various scoring functions. For example, one of the functions in our experiment calculates the token overlap-based groundedness score of utterance, u with respect to $\mathcal{K}$. These rewards are concatenated to every utterance of $\mathcal{D}$, producing the reward-augmented dialogue $\mathcal{D}'$.[4]

III. Feedback $\mathcal{D}'$ and $\mathcal{K}$ are fed into $\mathcal{M}$ to obtain natural language feedback $\mathcal{F}$ on improving the rewards. Following Madaan et al. (2023), we use the phrase 'actionable feedback' in the prompt to obtain a concrete plan to improve the scores.

IV. Refinement Finally, we feed the knowledge $\mathcal{K}$, the feedback $\mathcal{F}$, and the initial dialogue $\mathcal{D}$ to $\mathcal{M}$. The model is prompted to refine the dialogues based on the feedback provided.

3.2 Dataset Generation

We generate a dataset of meta-reviewing dialogues for fine-tuning dialogue agents. For this purpose, we test several models and perform automated and human evaluations to select the best configuration.

Models We use the following models namely, GPT 3.5 (dubbed 'ChatGPT'), LlaMa-2 13B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), and Mixtral (Jiang et al., 2024).[5]

Prompts ReMuSE relies on three prompt templates for the initial dialogue generation (I.), feedback (III.), and refinement (IV.) step. The prompts involve a step-specific instruction and include the knowledge source, i.e. the reviews for a paper, the paper's title and its publication type (long or short paper). We test multiple variations of the prompts to account for model sensitivity.[6].

Underlying Data To generate a dialogue, ReMuSE receives a paper and its reviews as the input. As a basis to generate a dataset, we rely on the ORSUM (Zeng et al., 2023) dataset consisting of reviews and meta-reviews for papers from 39 conferences collected using OpenReview.[7] We sample papers with exactly three reviews to fit most LLMs' context windows resulting in **3064** dialogues.

Reward Metrics We use three feedback metrics as rewards to assess the specificity and groundedness of generated dialogues. Knowledge Precision (K-Prec) measures token overlap between an utterance and its knowledge source (Adlakha et al., 2023). Q2-score (Honovich et al., 2021) evaluates faithfulness via question generation, answer generation, and natural language inference, leveraging pre-trained T5 (Raffel et al., 2020), ALBERT-Large (Lan et al., 2020), and DeBERTa-Large-MNLI (He et al., 2021) models respectively. Both capture dialogue groundedness. Specificity scores the level of detail in utterances using Specifiteller (Li and Nenkova, 2015).

Resulting Datasets Using each LLM, we generate synthetic dialogues. Due to the associated API costs for using closed LLMs, we follow previous works (Bian et al., 2024; Chen et al., 2024) that rely on a lower number of examples (e.g., 100) for evaluation. Consequently, we sample 200 examples to analyze the performance of ReMuSE and publicly release **3064** meta-reviewing dialogues generated with the best ReMuSE configuration.

3.3 Dataset Evaluation

We discuss various dialogue quality metrics employed in automated and human evaluation reflecting the desiderata for GROUNDEDNESS and SPECIFICITY. We compare the synthetically generated dialogues to a human-generated dialogue dataset that we use as evaluation data for comparison along the various evaluation setups.

3.3.1 Human Dialogue Dataset

We compare our synthetically generated dialogues with real human meta-reviewing decision-making dialogues using the *Dagstuhl* dataset (Balke et al., 2023). In this dataset, senior and junior scientists acted as meta-reviewers and assistants, respectively, in a wizard-of-oz setup, where the meta-reviewers used dialogue agents to gather information and decide on paper acceptance. The dataset is private and consists of 16 dialogues corresponding to 4 research papers.[8]

3.3.2 Automatic Evaluation

Setup We employ three measures (K-Prec, Q2-score, specificity), which we compute automati-

[4]The evaluator is completely independent of $\mathcal{M}$.

[5]More details about the models in §A.3.

[6]The prompts and further details are in §A.6

[7]There are no explicit licenses attached to OpenReview for third-party data processing. We use the dataset from the author's publicly available codebase.

[8]Additional details about the dataset is in §A.4

Model	Rwd.	K-Prec	Q2-F1	Q2-NL	Spec.
ChatGPT	-	$0.39_{.01}$	$0.15_{.01}$	$0.15_{.01}$	$0.45_{.13}$
LLaMa	-	$0.37_{.01}$	$0.15_{.13}$	$0.16_{.01}$	$0.46_{.01}$
Mistral	-	$0.38_{.01}$	$0.15_{.02}$	$0.16_{.01}$	$0.41_{.02}$
Mixtral	-	$0.38_{.01}$	$0.17_{.01}$	$0.19_{.02}$	$0.42_{.02}$
ChatGPT	ReMuSE	$0.74_{.12}$	$0.22_{.02}$	$0.23_{.18}$	$0.63_{.02}$
LLaMa	ReMuSE	$0.74_{.02}$	$0.23_{.02}$	$0.24_{.02}$	$0.63_{.02}$
Mistral	ReMuSE	$0.76_{.01}$	$0.30_{.02}$	$0.33_{.02}$	$0.65_{.22}$
Mixtral	ReMuSE	$\mathbf{0.78}_{.01}$	$\mathbf{0.32}_{.01}$	$\mathbf{0.34}_{.02}$	$\mathbf{0.72}_{.02}$

Table 1: Performance of LLMs before and after incorporating ReMuSE in terms of K-Precision (K-Prec), Q2-F1, Q2-NLI and Specificity (Spec). We average over three prompts and report the standard deviation.

Mode	Models	Coop.	Coh.	Eng.	Plau.
Pre-Refine.	ChatGPT	2.65	2.71	2.74	2.15
	LLaMa	2.53	2.13	2.11	2.45
	Mistral	2.34	2.22	2.18	2.19
	Mixtral	2.35	2.23	2.26	2.45
Post-Refine.	ChatGPT	3.01	**2.91**	**2.93**	2.75
	LLaMa	3.17	2.51	2.32	3.06
	Mistral	3.28	2.63	2.41	3.53
	Mixtral	**3.37**	2.79	2.85	**3.58**

Table 2: Human Evaluation of the generated dialogues before and after-refinement across different dimensions. All the scores are on a scale of 1 (worst) -4 (best).

cally given the knowledge source $\mathcal{K}$, and a dialogue utterance u, part of the generated dialog. These metrics are also employed as **rewards** within ReMuSE. The evaluation is performed on an utterance level. The intuition behind using the same measures for automated evaluation and for refinement is to understand whether ReMuSE can, in fact, guide the LLMs towards better generations reflected in exactly these dimensions we optimize for. We complement the automatic evaluation with our human evaluation to verify the improved quality.

For the final evaluation, we compute the K-Prec and Q2-scores only over the dialogue agent's utterances since we require only the dialogue agent to be faithful. The meta-reviewer (as an information seeker) need not fulfill such criteria since they can bring up new facts independent of the knowledge source at any point in time (Dziri et al., 2022). Nevertheless, all the utterances for the meta-reviewing dialogues are supposed to be highly specific. This stems from our observation that 65.24% utterances in the human dataset (cf. §3.2) have a score of 0.5 or above, indicating a significant presence of technical details. Hence, we measure the Specificity score over all utterances.

Overall results and comparison to human dialogues We start by comparing the performance for LLMs before and after incorporating ReMuSE (with all three rewards) in Table 1. We observe that ReMuSE consistently improves performances for all models across all dimensions, with Mixtral performing the best. We obtain maximum gains for the 'K-Prec' scores; for instance, Mixtral ReMuSE shows an improvement of 105% over its baseline performance. We achieve massive improvements in the Q2-F1 and Q2-NLI scores ($\sim$ 88% and 79% respectively for Mixtral). We also attain substantial improvement in specificity ($\sim$ 72% for Mixtral).[9][10] In order to compare human vs synthetic dialogues, we analyze K-Prec, Q2, and Specificity scores for all utterances in human and synthetic meta-reviewing dialogue datasets, along with {2,3,4}-gram token diversities ({n}-gram vocabulary size) for meta-reviewer utterances in Fig 2. LLM-based dialogue agents outperform humans in Q2 and K-Prec scores (cf. Figures 2c and 2d), demonstrating superior groundedness to the knowledge source. Human meta-reviewers exhibit lower groundedness, possibly due to off-topic divergence: for instance, in one of the dialogues, the meta-reviewer discusses datasets available for a related task not mentioned in the reviews. Regarding Specificity (Fig 2b), LLM-based agents perform comparable to humans, indicating their feasibility as meta-reviewing assistants. However, a significant gap exists between real and LLM meta-reviewers in terms of token diversity and specificity (cf. Fig 2a and 2b), suggesting LLMs still lack the expert knowledge for simulating real-world meta-reviewers.

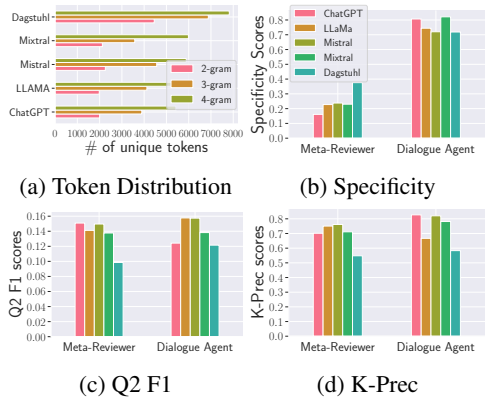
3.3.3 Human Evaluation

Setup Since using identical metrics for rewards and automated evaluation can be questioned for coincidental correlation, we further run a human evaluation. Following (Dziri et al., 2022; Wu et al., 2022b), we let humans evaluate the generated dialogs according to the following criteria: **Co-operativeness** evaluates the helpfulness of the dialogue agent's response to the query, **Coherence** specifies consistency of the overall dialogue, and **Engagingness** implies how engagingly the dialogue agent takes the discussion forward. Moreover, akin to faithfulness discussed before, we introduce a **Plausibility** score to measure the groundedness of the dialogue agent's responses to the paper's reviews. All the dialogues are rated on a Likert scale of $1-4$ and the ratings are averaged over each metric as done in prior works (Rashkin et al., 2021).[11]

[9] Table 6 shows the results for all models.

[10] Ablation study in §A.9 and cross-domain analysis in §A.8

[11] The instructions to the annotators are in §A.16



(a) Token Distribution (b) Specificity

(c) Q2 F1 (d) K-Prec

Figure 2: Comparison of utterances in human and synthetically generated dialogues in terms of (a) Token Distribution, (b) Specificity, (c) Q2 F1, and (d) K-Prec.

Overall results and comparison to human dialogues Four Ph.D. students with NLP backgrounds evaluated 70 dialogues, split between baseline LLMs (pre-refinement) and ReMuSE (post-refinement), using human evaluation metrics on a 1-4 Likert scale. Krippendorff's α for the dialogues showed substantial agreement: co-operativeness (0.68), coherence (0.71), engagingness (0.73), and plausibility (0.78). We observe that post-refinement, all models improved, with Mixtral excelling in co-operativeness and plausibility consistent with the automated evaluation. However, ChatGPT excels in coherence and engagingness likely due to its reported use of convincing language in responses (Lozić and Štular, 2023). We further conducted correlation analysis (cf. Fig 13,§A.19) of the automated and human evaluation metrics which revealed strong alignment between the metrics with similar objective (e.g., K-Prec and Plausibility), validating our evaluation approach.

Three Ph.D. students with an NLP background assessed the synthetic and human-generated dialogues on a Likert scale of $1-4$ as detailed in Sec §3.3.3. Kripendorff's α values for co-operativeness, coherence, engagingness, and plausibility were 0.68, 0.69, 0.70, and 0.74 respectively. Table 3 presents the resulting ratings. The synthetic dialogues were rated higher for co-operativeness, plausibility and, coherence. The lower ratings for the human dataset may be due to the setup where humans acted as dialogue agents, occasionally resulting in off-topic responses. For instance, in one of the dialogues, the dialogue agent discusses how the task in discussion is not equivalent to another task in the same domain, which is not grounded in the provided reviews. We also found occasional grammatical errors that may have also resulted in lower coherence scores, a common finding in previous works (e.g., Wu et al., 2022b). However, human dialogues were rated higher for engagingness due to the diverse meta-reviewer questions leading to longer and more detailed responses.

Setup	Coop.	Coh.	Eng.	Plau.
Dagstuhl	3.50	3.10	**3.10**	2.95
ReMuSE	**3.52**	**3.86**	3.00	**3.10**

Table 3: Human evaluation of dialogues from the human dialogue dataset (*Dagstuhl*) and our proposed ReMuSE. The evaluation scale ranges from 1 (worst) - 4 (best).

To reiterate, we evaluate the generated dialogues using automated and human assessments for different purposes. Automated metrics measure GROUNDEDNESS and SPECIFICITY against the knowledge source, while human evaluation focuses on flow and naturalness. Our results show LLM-generated dialogues are promising for meta-reviewing assistants with high faithfulness and specificity, but low token diversity in simulating meta-reviewer utterances indicates room for improvement. In Section §4, we discuss how finetuning can help address these limitations.

4 Dialogue Agents for Meta-Reviewing Assistance

With the synthetically generated dataset for the best-performing ReMuSE configuration, we now shift our focus to the task of finetuning dialogue agents for meta-reviewing assistance.[12]

4.1 Experimental Setup

Task Given a knowledge source $\mathcal{K}$ and a dialogue history $[u_1, u_2, \ldots, u_t]$, the task is to generate a response u_{t+1} grounded in $\mathcal{K}$. The title and reviews for every paper serve as the knowledge source.

Models Following Daheim et al. (2023), we select Flan-T5 (Longpre et al., 2023) (xl), T5 (Raffel et al., 2020) (3B), and OPT (Zhang et al., 2022) (2.7B) for experiments. For further comparison, we also prompt ChatGPT in a 0-shot setup and use the best model from the ReMuSE setup, Mixtral.[13]

Data and Evaluation We split the **3064** meta-reviewing dialogues having an average of **12.24** utterances per dialogue in our dataset into 60% for training, 20% for validation, and 20% for testing, respectively. In terms of evaluation, following Daheim et al. (2023), we use automated metrics - SacreBLEU (Post, 2018) and BERTScore (Zhang

[12] Analysis of the synthetic dataset is in §A.12

[13] The hyper-parameter details are provided in §A.13

Models	Fl.	Rel.	Faith.	Obj.	Help.
OPT	3.82	3.84	4.04	4.06	4.21
T5	3.91	3.98	4.12	4.21	4.32
Flan-T5	**4.21**	**4.18**	**4.58**	**4.62**	**4.71**
ChatGPT	3.84	3.82	3.28	3.91	3.85
ReMuSE	3.92	3.94	4.11	4.11	4.24

Table 4: Human evaluation of model responses on Fluency (Fl.), Relevance (Rel.), Faithfulness (Faith.), Objectivity (Obj.), and Helpfulness (Help.) scored from 1 (worst) to 5 (best).

et al., 2020) to measure response similarity to ground truth. We also report K-Prec and Q2 scores to assess faithfulness to $\mathcal{K}$. Additionally, following Rashkin et al. (2021), we perform human evaluation on fluency, relevance, faithfulness, and objectivity, plus a helpfulness score reflecting decision-making utility. All human ratings use a 1–5 scale, and the ratings are averaged over each metric.

4.2 Results and Discussion

Automated Evaluation We report model performances in Table 11 (§A.10). Flan-T5 achieves the highest K-Prec score (68.2), aligning with findings from prior work (Daheim et al., 2023). In contrast, ChatGPT shows significantly lower performance (K-Prec 42.1), underscoring the importance of high-quality task-specific supervision. ReMuSE demonstrates strong faithfulness, as reflected by its high K-Prec (67.6) scores. However, its BLEU and BERT scores are notably lower than those of fine-tuned models. This discrepancy stems from ReMuSE being optimized for faithful and diverse generation rather than surface-level alignment with ground truth due to its inference-only setup. In contrast, fine-tuned models benefit from task-specific supervision, leading to better alignment with reference responses. (Hsieh et al., 2023; Fu et al., 2024).

Human Evaluation Following Santu et al. (2024), we evaluate dialogue agents on their ability to assist meta-reviewers by rating responses to key queries on contributions, strengths, weaknesses, improvements, and literature review. Following Purkayastha et al. (2023), we ask two senior NLP Ph.D. students with over three years of reviewing experience to rate 60 responses per model on a 1–5 Likert scale.[14] Krippendorff's α for fluency, relevance, faithfulness, objectivity, and helpfulness are 0.58, 0.64, 0.69, 0.66, and 0.74, respectively. The scores are substantial given the subjectivity of the task (Kennard et al., 2022).[15] We show the results in Table 4. Flan-T5 scores highest, aligning with automated evaluations, while ChatGPT scores lowest on faithfulness and objectivity. ReMUSE outperforms ChatGPT but remains behind fine-tuned models, likely due to its diversity-driven objectives and verbose responses. Importantly, the responses from ReMuSE end with follow-up questions that stem from the pre-training objectives in LLMs, which favor full dialogue generation. This limits their ability to generate focused, knowledge-grounded responses based on provided dialogue history as witnessed in prior work (Chen et al., 2025). Helpfulness correlates strongly with faithfulness and objectivity (Fig. 8), explaining ChatGPT's lower helpfulness. These results support the quality of our synthetic training data.

5 Deploying Meta-Reviewing Dialogue Agents: A Within-subject Study

We analyze the trained dialogue agent's effectiveness in real-world meta-reviewing, focusing on time and meta-review quality.

Setup We conduct a within-subject controlled experiment where participants write meta-reviews with and without dialogue agent support.[16] Following prior work on user studies in peer-reviewing (Santu et al., 2024), we ask ten Ph.D. students with peer-reviewing experience to evaluate three papers each under both conditions (30 papers total).[17] Following Shen et al. (2022), we assess meta-review quality using (i) *Content Relevance* (alignment with review aspects, rated 1-4), (ii) *Decision Correctness* (alignment with the gold decision, binary 0/1), and (iii) *Coverage* (review diversity, rated 1-4). We explain the conference-specific guidelines to the participants in order to ensure consistency with the ground-truth meta-reviews.

Results As shown in Table 13 (§A.21), meta-reviews generated with the dialogue agent outperform those written without it in Content Relevance (3.87 vs. 3.42) and Coverage (3.76 vs.3.21), with a slight improvement in Decision Correctness (0.72 vs.0.60). Interaction with the agent takes **20 minutes** on average, compared to **35 minutes** while writing after reading the reviews, offering a more efficient reviewing process without compromising reviewer autonomy. We further analyze the 30 papers — 15 borderline, 7 accepted, and 8 rejected. We observe that the borderline cases produce longer dialogues (12.2 turns) than accepted (8.3) or

[14] See §A.17 for instructions to annotators.

[15] We analyze the disagreements of the annotators in §A.20

[16] The interaction interface is in Fig 7, §A.21

[17] The participant details is in §A.18

rejected (6.1) ones, indicating greater deliberation. While the discussions on accepted/rejected papers focus on content and related work, borderline cases emphasize reviewer scores and confidence. We find that the agent provides clarification when prompted but refrains from final decisions. While accepted and rejected outcomes generally align with ground truth, borderline cases show variability, likely due to external factors such as acceptance rates.

Case Study with Non-Experts While meta-reviewing is typically done by experienced researchers in the field, we also explore if individuals from interdisciplinary backgrounds can effectively use the tool. Five PhD students from non-CS fields, with no prior peer-review experience, wrote meta-reviews for 3 papers each under both setups.[18] Results in Table 13 (§A.21) show that reviews with the dialogue agent scored higher in Content Relevance (3.65 vs 3.10), Coverage (3.50 vs 2.95), and Decision Correctness (0.60 vs 0.58). Participants spent an average of **35 minutes** using the agent versus **80 minutes** writing without it. This demonstrates the agent's promise in helping newcomers learn meta-reviewing efficiently.

6 Related Work

Meta-Review Generation Meta-reviewing is commonly framed as multi-document summarization on the reviews. Wu et al. (2022a) treat reviews, author responses, and reviewer discussions as an argumentative graph to generate comprehensive meta-reviews. Li et al. (2023a) encode entire reviewer conversation threads using a multi-task approach to meta-review generation. Shen et al. (2022) introduce control codes for generating diverse meta-reviews. Zeng et al. (2023) use LLMs in a self-refinement paradigm based on task-specific checklists. Recently, Santu et al. (2024) analyzed the performance of LLMs for generating various meta-reviewing aspects using a prompting taxonomy. In this work, we redefine meta-reviewing as a dialogue to aid meta-reviewers in analyzing the reviews for effective decision making rather than automatically generating the final text – this way, we respect and acknowledge the scientific autonomy of the meta-reviewers as discussed in Section §2.

AI-assisted decision making In this work, we frame meta-reviewing as a decision-making process. The role of AI in supporting human decisions has expanded across fields like finance, law, and technology (Lee et al., 2020; Lai et al., 2023), following the collaborative decision-making paradigm (Leitão et al., 2022; Lin et al., 2023). While Lin et al. (2023) explore reward-based models with fixed solutions, we focus on meta-reviewing, where no single best solution exists, and decisions depend on the meta-reviewer's expertise.

Synthetic Dialogue Generation We employ synthetically generated dialogues to finetune dialogue agents for assisting meta-reviewers. These dialogues are knowledge-intensive and require utterances to be anchored in the underlying documents (reviews). Prior studies highlight the issue of limited training data in this context, leading to approaches for synthetic data generation (Dai et al., 2022; Bao et al., 2023). These efforts have mainly explored fine-tuning for dialogue generation on structured knowledge sources. In contrast, we harness the potential of LLMs as cost-effective data curators for knowledge-intensive dialogue generation. Though there have been studies prompting LLMs for related purposes (Møller et al., 2023; Chen et al., 2023a), we are the first to explore the generation of full-length knowledge-grounded decision making dialogues. This framing introduces multiple additional constraints on the quality of the output, like high faithfulness and technicality of the utterances. Instead of improving a single constraint only or assuming the availability of ground truth feedback (Chen et al., 2023b; Xu et al., 2024), we fulfill our requirements through a novel zero-shot self-refinement strategy (Madaan et al., 2023), which enhances the generations in multiple measurable aspects.

7 Conclusion

In this work, we explored the potential of dialogue agents to assist meta-reviewers in the decision-making process, offering an alternative to traditional summarization approaches. We addressed the data scarcity challenge with a self-refinement method for generating diverse, faithful decision-making dialogues using LLMs. We found that models fine-tuned on our data significantly outperformed their zero-shot counterparts. Using a controlled experiment, we demonstrated that dialogue agents reduced the meta-reviewing time by almost half while maintaining reviewer autonomy. Our findings highlight the potential of AI-driven tools to improve peer-reviewing efficiency and encourage further research in this area.

[18] Participant details in §A.18

Limitations

In this work, we propose a new perspective towards meta-reviewing and introduce a novel method along with a dataset to assist meta-reviewers and other decision-makers in relevant domains. However, our work comes with several limitations. Firstly, we employ both closed and open-source large language models that are standard for dialogue generation. These models have been criticized for producing harmful or biased content. We do not address biases such as socio-demographic factors, given the assumed neutrality of the decision-making domain. Secondly, we do not consider potential bias in the knowledge sources, such as paper reviews, which may include subjective content. Our focus is on generating faithful and diverse dialogues; analyzing component-wise bias is left to future work. Thirdly, regarding transparency, AI models often lack interpretability, making it difficult to trace how recommendations are generated. We do not aim to improve interpretability in this work, and the responsibility for validating suggestions lies with the meta-reviewer. Accountability is especially important when agent outputs influence decisions. While our agent is designed not to make final judgments, it may occasionally offer flawed suggestions. To mitigate this, reviewer training and clear usage guidelines are essential. Lastly, our study centers on English-language reviews in AI conferences. Future work could extend this to multilingual settings and other domains, such as the humanities. We also emphasize that the agent cannot compensate for poor-quality reviews—addressing review quality remains a critical, orthogonal research direction.

Ethics Statement

In this work, we provide a detailed study of the dialogue generation capabilities of Large language models for meta-reviewing. Consequently, this work has an ethical dimension since it can be perceived as an attempt to replace human meta-reviewers with LLMs. We would like to emphasize that we solely focus on assisting the meta-reviewers in decision making using dialogue agents rather than providing them with any subjective opinions. Unlike previous work, we do not attempt in any way to generate meta-reviews and strongly believe that such a task is a **'Human-In-The-Loop'** process which requires highly skilled researchers (meta-reviewers) operating as humans to make the loop effective.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *ArXiv preprint*, abs/2307.16877.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.

Wolf-Tilo Balke, Andreas Vlachos, Davide Ceolin, Milad Alshomary, Nils Dycke, Sukannya Purkayastha, Iryna Gurevych, Anne Lauscher, and Tilman Beck. 2023. 5.4 scholarly argumentation as a community dialogue. *Towards a Unified Model of Scholarly Argumentation*, page 202.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *ArXiv preprint*, abs/2303.16421.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *ArXiv preprint*, abs/2309.13007.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023a. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023b. Iterative translation refinement with large language models. *ArXiv preprint*, abs/2306.03856.

Xiangyan Chen, Yujian Gan, and Matthew Purver. 2025. Improving factuality for dialogue response generation via graph-based knowledge augmentation.

Xieling Chen, Haoran Xie, and Xiaohui Tao. 2022. Vision, status, and research topics of natural language processing. *Natural Language Processing Journal*, 1:100001.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation. *ArXiv preprint*, abs/2303.17574.

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

Tjibbe Donker. 2023. The dangers of using large language models for peer review. *The Lancet Infectious Diseases*, 23(7):781.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Youmna Farag, Charlotte Brand, Jacopo Amidei, Paul Piwek, Tom Stafford, Svetlana Stoyanchev, and Andreas Vlachos. 2022. Opening up minds with argumentative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4569–4582, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394, Mexico City, Mexico. Association for Computational Linguistics.

Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying helpful sentences in product reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 678–691, Online. Association for Computational Linguistics.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics*, 83(5):368–375.

Roberta Michnick Golinkoff, Erika Hoff, Meredith L Rowe, Catherine S Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child development*, 90(3):985–992.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv preprint*, abs/2006.03654.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv preprint*, abs/2401.04088.

Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *Journal of Medical Internet Research*, 26:e59439.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. 2023. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*, 5(11).

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. 2024. What can natural language processing do for peer review? *ArXiv preprint*, abs/2405.06563.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1369–1385, New York, NY, USA. Association for Computing Machinery.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Diogo Leitão, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. 2022. Human-ai collaboration in decision-making: beyond learning to defer. *ArXiv preprint*, abs/2206.13202.

Rosemary Lever and Monique Sénéchal. 2011. Discussing stories: On how a dialogic reading intervention improves kindergartners' oral narrative construction. *Journal of experimental child psychology*, 108(1):1–24.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2281–2287. AAAI Press.

Miao Li, Eduard Hovy, and Jey Lau. 2023a. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

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

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2023. Decision-oriented dialogue for human-ai collaboration. *ArXiv preprint*, abs/2305.20076.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Edisa Lozić and Benjamin Štular. 2023. Fluent but not factual: A comparative analysis of chatgpt and other ai chatbots' proficiency and originality in scientific writing for humanities. *Future Internet*, 15(10).

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv preprint*, abs/2303.17651.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *ArXiv preprint*, abs/2304.13861.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. 2023. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718, Online. Association for Computational Linguistics.

Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, and Matthew C. Williams Jr au2. 2024. Prompting llms to compose meta-review drafts from peer-review narratives of scholarly manuscripts. *ArXiv preprint*, abs/2402.15589.

Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the influence of explainable ai on automation bias. *arXiv preprint arXiv:2204.08859*.

Laurie A Schintler, Connie L McNeely, and James Witte. 2023. A critical examination of the ethics of ai-mediated peer review. *arXiv preprint arXiv:2309.12356*.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.

Daniel Smolyak, Margrét V Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. 2024. Large language models and synthetic health data: progress and prospects. *JAMIA Open*, 7(4):ooae114.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Ziqi Wang, Le Hou, Tianjian Lu, Yuexin Wu, Yunxuan Li, Hongkun Yu, and Heng Ji. 2023. Enable language models to implicitly learn self-improvement from data. *ArXiv preprint*, abs/2310.00898.

Mark Ware and Michael Mabe. 2009. An overview of scientific and scholarly journal publishing. *The STM report*, 1082:1083.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022a. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2189–2198, New York, NY, USA. Association for Computing Machinery.

Qingyang Wu, Song Feng, Derek Chen, Sachindra Joshi, Luis Lastras, and Zhou Yu. 2022b. DG2: Data augmentation through document grounded dialogue generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 204–216, Edinburgh, UK. Association for Computational Linguistics.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback. *ArXiv preprint*, abs/2311.09336.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2023. Scientific opinion summarization: Meta-review generation with checklist-guided iterative introspection. *ArXiv preprint*, abs/2305.14647.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

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

A Appendix

A.1 Differences to prior approaches

We highlight the differences of our proposed approach to prior meta-reviewing systems in Table 5.

A.2 ReMuSE Figure

We provide a detailed figure depicting every stage of our method, ReMuSE (Reward-based Multi-aspect Self Editing) in Fig 3.

A.3 Model Details

GPT 3.5 We use the *turbo* version of GPT 3.5 which is the chat-tuned variant (dubbed *'ChatGPT'*) of the Instruct-GPT class of models from OpenAI (Ouyang et al., 2022).[19] This model was trained with internet data and finetuned using reinforcement learning from human feedback (Bai et al., 2022). The training and dataset details are not publicly available. We use 'gpt35-turbo-0301'.

LLaMa-2 It is the upgraded version of the LLaMa family of models (Touvron et al., 2023) featuring a 40% increase in pre-training models' data and doubled context length compared to its predecessor. We employ the 13B-chat version in our experiments.[20]

Mistral This model is trained on the grouped query and sliding window attention that helps focus on words even outside the context window (Jiang et al., 2023). We use the 7B version.[21]

Mixtral It is an open-source LLM based on a mixture of experts (Jiang et al., 2024). Each feed-forward block has access to 8 parameter groups and utilizes two of them to process a token.[22]

We use the Azure OpenAI service to prompt ChatGPT. The rest of the models are implemented using huggingface transformers (Wolf et al., 2020). Following previous work on data generation with LLMs (Adlakha et al., 2023), we use a high temperature of 0.95 and employ top-p sampling (Holtzman et al., 2020) of 0.95 to avoid sampling repetitions.

[19] https://chat.openai.com/

[20] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[21] https://huggingface.co/mistralai/Mistral-7B-v0.1

[22] https://huggingface.co/mistralai/Mixtral-8x7B-v0.1

A.4 Dagstuhl dataset

For the comparison of our synthetically generated dialogues with real human meta-reviewing decision making dialogues we employ the *Dagstuhl* dataset (Balke et al., 2023) kindly provided by the authors. For this dataset, 4 senior and 4 junior scientists acted as meta-reviewers and assistants, respectively, in a wizard-of-oz scenario. The study was conducted on 4 papers selected from the NLPeer corpus (Dycke et al., 2023). The assistants were provided with the paper and reviews in advance, whereas the meta-reviewers were instructed to skim the reviews within 10 minutes. The meta-reviewers were asked to converse with the dialogue agents to gather information and eventually decide on the acceptance of the paper. The paper acceptance decisions for the dataset are displayed in Figure 4. Across the 4 papers, only 1 paper had a unanimous reject decision, while all the other papers received diverse verdicts.

A.5 Computational Budget

We ran all the experiments on Nvidia A100 80GB GPUs. None of the experiments consumed more than 36 hours.

A.6 Prompts for all the experiments

Our proposed method, ReMuSE, uses different prompting templates for the Initial Dialogue Generation, Feedback, and Refinement steps. The templates consist of an instruction specific to the particular ReMuSE step combined with our knowledge source, i.e., the paper's reviews, title, and type (long or short). We use multiple prompts for the initial dialogue generation to account for model sensitivity. We name these prompts as follows: extensive prompt (simulates the study in Balke et al. (2023)), paraphrased prompt (paraphrase of the extensive prompt), and tl:dr prompt (3-sentence summary of the extensive prompt). We provide full templates for these prompts along with the ones used for Feedback and Refinement in Appendix §A.6.1, §A.6.2 and §A.6.3, respectively.

A.6.1 Initial Dialogue Generation Prompts for all the experiments

We show the different prompts for initial dialogue generation below:

Aspect	Prior Methods (Summarization)	Our Approach (Decision-Making Dialogue)
Focus of Task	Frames meta-reviewing as summarization—extracts key points from reviews.	Frames it as a decision-making task with contextual human judgment.
Nature of Process	One-shot summarization leading to final decisions.	Interactive process that supports iterative reasoning.
Model Output	Produces summaries without adjusting for bias or reasoning depth.	Generates context-aware, review-grounded responses.
Decision Support	Offers limited guidance; summaries may lack depth.	Actively aids reviewers in making informed, reasoned decisions.

Table 5: Comparison between summarization-based and decision-making dialogue-based approaches for meta-reviewing.

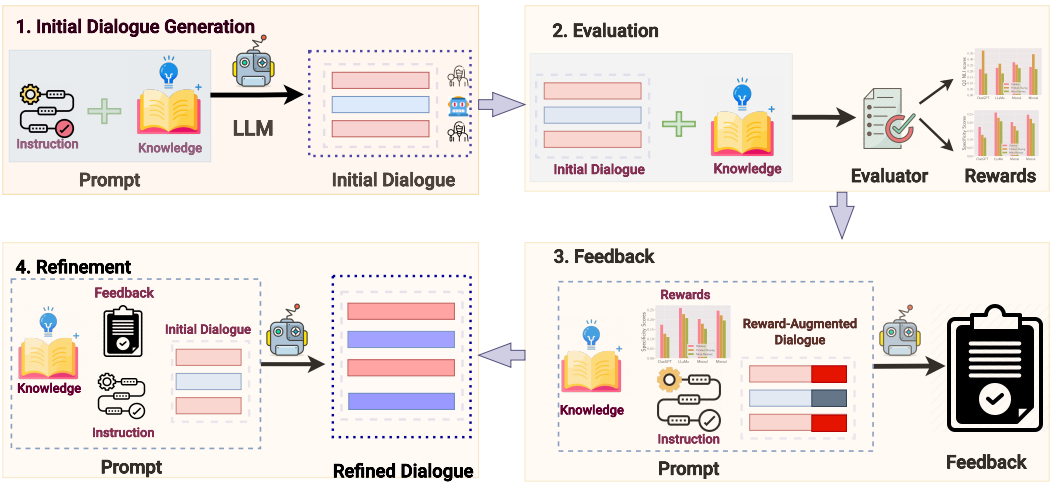


Figure 3: Overview of our **Re**ward-based **Mu**lti-aspect **S**elf **E**diting (ReMuSE) method. ReMuSE consists of four steps: 1. **Initial Dialogue Generation** in which we prompt an LLM with relevant documents (paper reviews) and instructions, 2. **Evaluation** of the dialogues by computing one or multiple measures (rewards), 3. Natural language **Feedback** Generation based on the computed rewards, 4. Self-**Refinement** of the dialogues based on the feedback.

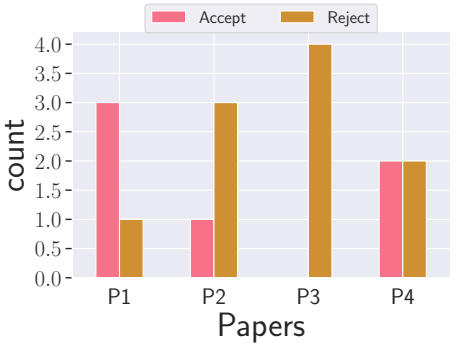


Figure 4: Diversity of Paper Acceptance Decisions in the Dagstuhl dataset

Extensive Prompt

Extensive Prompt (Meta-Reviewing)

Instruction: Generate a multi-turn dialogue between a meta-reviewer and a dialogue agent for reviews about a paper. Every turn should alternate between “Dialogue Agent" and “Meta Reviewer" utterances. The meta-reviewer must explicitly tell whether the paper is accepted or rejected at the nal turn. The dialogue ends with the meta-reviewer saying whether they accept or reject the paper. You need to simulate both the meta-reviewer and the dialogue agent. The dialogue agent has access to only the reviews and type of the paper. A meta-reviewer mainly asks questions and the dialogue agent should just answer from the reviews. A dialogue agent should never recommend anything/ give any opinions/ decide anything for the paper. A dialogue agent has no conscience of its own. A meta-reviewer can also ask follow-up questions and grill the dialogue agent for more information on the reviews. As a meta- reviewer, you should also weigh the importance of the condence of the reviewers while making a decision.

Title: [title]

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: ..,* ***Review 3****: ...*

Extensive Prompt (Product-Buying)

Instruction: Generate a multi-turn dialogue between a buyer who wants to buy a product and a dialogue agent for reviews about that product. Every turn should start with either "Dialogue Agent" followed by their utterance or "Buyer" followed by their utterance. The dialogue agent should always remain neutral and take a neutral stand in any case. The buyer should reach a decision to buy/not buy the paper at the final turn. You need to simulate both the buyer and the dialogue agent. The buyer does not have access to the reviews but the dialogue agent has access to the reviews and nothing else. It should answer any question about the reviews. A buyer mainly asks questions and the dialogue agent should refrain from offering advice if its opinion is sought. The dialogue agent should not recommend/advise anything regarding the product.

Title: “OP/TECH USA 1504372 Pro Loop Strap (Royal)".

Knowledge Source: *This strap is just what I needed for lugging my D700 + battery pack + It is comfortable and can handle a dslr and heavy lens. I bought this for the Nikon D700 camera and replaced the factory Nikon strap which is thinner. Very nice product with great adaptability and not made in China!...*

Extensive Prompt (Debates)

Instruction: Generate a multi-turn dialogue between a debate decision maker who needs to take a decision about which side wins a debate and a dialogue agent that has access to the arguments put forward by both the sides. Every turn should start with either "Dialogue Agent" followed by their utterance or "Decision Maker" followed by their utterance. You need to simulate both the decision maker and the dialogue agent. The decision maker does not have access to the arguments put forward by both the sides. The decision maker should just rely on the dialogue agent to know about the arguments from both the sides. A decision maker mainly asks questions and the dialogue agent answers. The dialogue agent should refrain from answering if its opinion/decision is sought. The dialogue agent has access to only the arguments of the debate. The dialogue agent should never decide on who wins the debate and should always take a neutral stand when any opinion is sought.

Title: “Abolish The Minimum Wage".

Knowledge Source: ***For arguments****: ’Let’s talk about what would happen to real people in real families ...’,* ***Against Arguments****: ‘We need to abolish the minimum wage. And there’s only one argument that matters on this issue, and it’s the moral argument. Does the minimum wage make the world a better...’*

Paraphrased Prompt

Paraphrased Prompt (Meta-Reviewing)

Instruction: Create a multi-turn dialogue featuring a meta-reviewer and a dialogue agent discussing reviews of a paper. Each turn should begin with either "Dialogue Agent" or "Meta Reviewer" followed by their respective statements. The meta-reviewer must make a final decision to "accept" or "reject" the paper and explicitly state their choice at the end. The dialogue concludes with the meta-reviewer indicating whether they accept or reject the paper. The meta-reviewer, aware only of the paper’s title, lacks access to reviews or information about the paper type and relies on the dialogue agent for this data. The dialogue agent has access solely to reviews and the paper type, responding neutrally and abstaining from offering opinions, recommendations, or decisions. If asked for its opinion, the dialogue agent must respond with "I do not know." The meta-reviewer primarily poses questions, and the dialogue agent answers based on the reviews, with the option to inquire about reviewer condence. The meta-reviewer may ask follow-up questions and press the dialogue agent for additional details from the reviews. It’s important to note that the dialogue agent has no inherent conscience or personal opinions.

Title: [title]

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: ..,* ***Review 3****: ...*

Paraphrased Prompt (Product-Buying)

Instruction: Create a multi-turn conversation between a prospective buyer interested in purchasing a specic product and a dialogue agent knowledgeable about reviews for that product. Each turn should commence with either "Dialogue Agent" followed by their statement or "Buyer" followed by theirs. The dialogue agent must maintain a neutral stance throughout and respond solely based on the reviews available. The buyer, who has not read the reviews, will make a denitive decision on whether to purchase the product or not in the nal turn. The dialogue agent should refrain from expressing opinions and respond with 'I do not know' if asked for personal input. It's essential to clarify that the dialogue agent is not a seller of the product and possesses information solely from the reviews.

Title: "OP/TECH USA 1504372 Pro Loop Strap (Royal)".

Knowledge Source: *This strap is just what I needed for lugging my D700 + battery pack + It is comfortable and can handle a dslr and heavy lens. I bought this for the Nikon D700 camera and replaced the factory Nikon strap which is thinner. Very nice product with great adaptability and not made in China!...*

Paraphrased Prompt (Debates)

Instruction: Craft a multi-turn conversation involving a decision maker tasked with determining the winning side in a debate, and a dialogue agent equipped with information on the arguments presented by both sides. Each turn should initiate with either "Dialogue Agent" or "Decision Maker" followed by their respective statements. The decision maker, who is unaware of the arguments, must make a conclusive decision on the debate's victor in the nal turn. The dialogue agent should remain impartial, responding only based on the arguments, and stating "I do not know" if asked for its opinion. The decision maker relies solely on the dialogue agent for information about the arguments from both sides. The dialogue agent possesses access solely to the arguments and is prohibited from making decisions or expressing opinions on the debate outcome. **Title**: "Abolish The Minimum Wage".

Knowledge Source: ***For arguments****: 'Let's talk about what would happen to real people in real families if the U.S. were to go down this untested and quite radical path of eliminating the minimum wage as our opponents urge...', **Against Arguments**: 'We need to abolish the minimum wage. And there's only one argument that matters on this issue, and it's the moral argument. Does the minimum wage make the world a better place?...'*

TL:DR **Prompt**

TL:DR Prompt for Initial Dialogue Generation (Meta-Reviewing)

Instruction: Generate a multi-turn dialogue between a meta-reviewer and a dialogue agent regarding a paper's reviews, with the meta-reviewer aiming to decide whether to accept or reject the paper. The dialogue agent, relying on reviews and aware of the paper type, must respond to the meta-reviewer's questions and refrain from providing opinions or recommendations, while the meta-reviewer weighs the importance of reviewer confidence in making the final decision

Title: [title].

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ..., **Review 2**: .., **Review 3**: ...*

TL:DR Prompt for Initial Dialogue Generation (Product-Buying)

Instruction: Generate a multi-turn dialogue between a buyer and a neutral dialogue agent regarding a product, with the buyer making a nal decision to buy or not. The dialogue agent, equipped with access to reviews, answers the buyer's questions, refraining from recommendations or opinions, while the buyer, without access to reviews, seeks information to reach a purchasing decision.

Title: "OP/TECH USA 1504372 Pro Loop Strap (Royal)".

Knowledge Source: *This strap is just what I needed for lugging my D700 + battery pack + It is comfortable and can handle a dslr and heavy lens. I bought this for the Nikon D700 camera and replaced the factory Nikon strap which is thinner. Very nice product with great adaptability and not made in China!...*

TL:DR Prompt for Initial Dialogue Generation (Debates)

Instruction: Simulate a multi-turn dialogue between a decision maker and a dialogue agent for a debate, where the decision maker, aware only of the debate's topic, seeks information from the dialogue agent, who has access to arguments from both sides. The decision maker aims to reach a conclusion about which side wins, while the dialogue agent remains neutral, providing answers solely based on the arguments presented, refraining from expressing opinions or making decisions

Title: "Abolish The Minimum Wage".

Knowledge Source: ***For arguments****: 'Let's talk about what would happen to real people in real families if the U.S. were to go down this untested and quite radical path of eliminating the minimum wage as our opponents urge...', **Against Arguments**: 'We need to abolish the minimum wage. And there's only one argument that matters on this issue, and it's the moral argument. Does the minimum wage make the world a better place?...'*

A.6.2 Prompt Templates for ReMuSE feedback

We have different versions of the same prompt used for multiple metrics. The prompt for refinement using all the metrics is provided below:

Prompt for ReMuSE feedback (Meta-Reviewing)

Instruction: Given the knowledge source and the dialogue, please provide actionable feedback to improve the dialogues. The feedback should just be for the overall dialogue and should start with 'Feedback:'. A Q2 F1 score, Q2 NLI score, KPrecision, and specificity scores follow each utterance. The Q2 and Kprecision scores measure the groundedness of the response to the provided knowledge source. The Specificity scores measure the technicality of the response. The feedback should try to improve all scores for the dialogue agent and the Specificity scores for the meta-reviewer.

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: ..,* ***Review 3****: ...*

Dialogue: *Meta-Reviewer: Hello Dialogue Agent. Can you tell me more about this paper?, F1: 0.0, NLI: 0.0, Kprec: 0.0, Specificity: 0.1 \n Dialogue Agent: Ofcourse! This is a paper about CodexGLUE, a machine learning benchmark dataset for code understanding and generation, F1: 0.12, NLI: 0.34, Kprec: 0.45, Specificity: 0.7 \n MetaReviewer: ...*

Prompt for ReMuSE feedback (Debates)

Instruction: Given the knowledge source and the dialogue, please provide actionable feedback to improve the dialogues. The feedback should just be for the overall dialogue and should start with 'Feedback:'. A Q2 F1 score, Q2 NLI score, KPrecision, and specificity scores follow each utterance. The Q2 and Kprecision scores measure the groundedness of the response to the provided knowledge source. The Specificity scores measure the technicality of the response. The feedback should try to improve all scores for the dialogue agent and the Specificity scores for the decision-maker.

Title: "Abolish The Minimum Wage".

Knowledge Source: ***For arguments****: 'Let's talk about what would happen to real people in real families if the U.S. were to go down this untested and quite radical path of eliminating the minimum wage as our opponents urge...',* ***Against Arguments****: 'We need to abolish the minimum wage. And there's only one argument that matters on this issue, and it's the moral argument. Does the minimum wage make the world a better place?...'*

Dialogue: *Decision Maker: Hello, can you tell the arguments for and against debating for minimum wages? F1: 0.2, NLI: 0.2, KPrec:0.01, Specificity: 0.2 \n Dialogue Agent: The debate over abolishing minimum wage centers on two perspectives... F1: 0.4, NLI: 0.39, KPrec: 0.45, Specificity: 0.6 ...*

Prompt for ReMuSE feedback (Product-Buying)

Instruction: Given the knowledge source and the dialogue, please provide actionable feedback to improve the dialogues. The feedback should just be for the overall dialogue and should start with 'Feedback:'. A Q2 F1 score, Q2 NLI score, KPrecision, and specificity scores follow each utterance. The Q2 and Kprecision scores measure the groundedness of the response to the provided knowledge source. The Specificity scores measure the technicality of the response. The feedback should try to improve all scores for the dialogue agent and the Specificity scores for the buyer.

Knowledge Source: *This strap is just what I needed for lugging my D700 + battery pack + It is comfortable and can handle a dslr and heavy lens. I bought this for the Nikon D700 camera and replaced the factory Nikon strap which is thinner...*

Title: "OP/TECH USA 1504372 Pro Loop Strap (Royal)".

Dialogue: *Buyer: Hi there, I am interested in buying the OP/TECH USA 1504372 Pro Loop Strap (Royal). can you tell me more about it? F1: 0.0, NLI: 0.0 \n Dialogue Agent: Sure! What would you like to know? F1: 0.0, NLI: 0.0, Kprec: 0.0, Spec: 0.1 \n Buyer: How comfortable are these for long travel hours? F1: 0.1, NLI: 0.1, Kprec: 0.05, Specificity: 0.4 \n Dialogue Agent: According to the reviews, its comfortable to wear for long hours, F1: 0.2, NLI: 0.2, Kprec: 0.25, Specificity: 0.5 \n Buyer: ...*

A.6.3 Prompt Templates for ReMuSE refinement

We provide the refinement prompts here:

Prompt for Refinement (Meta-Reviewing)

Instruction: Given the feedback, knowledge source, and dialogue improve the dialogue based on the feedback provided. The output should just be the new dialogue.

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: ..,* ***Review 3****: ...*

Feedback: *It would be helpful to provide a clearer explanation of the tasks' relevance to software development in response to reviewer one's concerns. Additionally, it may be useful to mention that the paper discusses the effectiveness of the baselines and provides an ablation study. Finally, it would be good to acknowledge reviewer three's comment explicitly and mention any plans to revise the descriptions of the datasets.*

Dialogue: *Meta-Reviewer: Hello Dialogue Agent. Can you tell me more about this paper? \n Dialogue Agent: Ofcourse! This is a paper about CodexGLUE, a machine learning benchmark dataset for code understanding and generation, \n MetaReviewer: Interesting. Can you give me some more information about the reviews that have been given for this paper? \n Dialogue Agent: ...*

Prompt for Refinement (Product-Buying)

Instruction: Given the feedback, knowledge source, and dialogue, improve the dialogue based on the feedback provided. The output should just be the new dialogue.

Knowledge Source: *This strap is just what I needed for lugging my D700 + battery pack + It is comfortable and can handle a dslr and heavy lens. I bought this for the Nikon D700 camera and replaced the factory Nikon strap which is thinner. Very nice product with great adaptability and not made in China!...*

Feedback: *Here are some specific feedback points: 1. The buyer is hesitant about the price, and the dialogue agent could provide more information about the product's value and benefits to address these concerns. 2. The buyer asks about the reviews and ratings, and the dialogue agent could provide more specific examples or testimonials to showcase the product's positive feedback. 3. The dialogue agent could provide more information about the return policy and the retailer's customer service to alleviate any concerns the buyer may have.*

Dialogue: *Buyer: Hi there, I am interested in buying the OP/TECH USA 1504372 Pro Loop Strap (Royal). can you tell me more about it? \n Dialogue Agent: Sure! What would you like to know? \n Buyer: How comfortable are these for long travel hours? \n Dialogue Agent: According to the reviews, its comfortable to wear for long hours, \n Buyer: ...*

Domain	Model	Rewards	K-Prec	Q2-F1	Q2-NLI	Spec.
Meta-Reviewing	ChatGPT	-	0.387 ± 0.012	0.147 ± 0.011	0.152 ± 0.011	0.451 ± 0.127
	LLaMa	-	0.374 ± 0.014	0.145 ± 0.130	0.155 ± 0.014	0.461 ± 0.081
	Mistral	-	0.377 ± 0.012	0.151 ± 0.022	0.161 ± 0.014	0.407 ± 0.015
	Mixtral	-	0.378 ± 0.014	0.171 ± 0.014	0.188 ± 0.021	0.422 ± 0.021
	ChatGPT	K-Prec	0.679 ± 0.022	0.177 ± 0.121	0.197 ± 0.025	0.552 ± 0.034
	LLaMa	K-Prec	0.633 ± 0.016	0.162 ± 0.141	0.163 ± 0.012	0.563 ± 0.021
	Mistral	K-Prec	0.681 ± 0.014	0.242 ± 0.014	0.231 ± 0.021	0.611 ± 0.024
	Mixtral	K-Prec	0.702 ± 0.011	0.268 ± 0.021	0.263 ± 0.015	0.646 ± 0.026
	ChatGPT	Q2 (F1 + NLI)	0.689 ± 0.013	0.171 ± 0.011	0.173 ± 0.013	0.648 ± 0.018
	LLaMa	Q2 (F1 + NLI)	0.646 ± 0.017	0.178 ± 0.029	0.189 ± 0.013	0.571 ± 0.026
	Mistral	Q2 (F1 + NLI)	0.706 ± 0.017	0.268 ± 0.034	0.272 ± 0.012	0.626 ± 0.013
	Mixtral	Q2 (F1 + NLI)	0.721 ± 0.014	0.271 ± 0.016	0.265 ± 0.012	0.701 ± 0.028
	ChatGPT	Spec.	0.692 ± 0.015	0.180 ± 0.023	0.181 ± 0.019	0.652 ± 0.032
	LLaMa	Spec.	0.681 ± 0.010	0.183 ± 0.021	0.191 ± 0.016	0.592 ± 0.035
	Mistral	Spec.	0.703 ± 0.021	0.292 ± 0.021	0.300 ± 0.021	0.639 ± 0.020
	Mixtral	Spec.	0.704 ± 0.012	0.285 ± 0.017	0.269 ± 0.014	0.689 ± 0.020
	ChatGPT	K-Prec + Q2	0.701 ± 0.021	0.178 ± 0.014	0.188 ± 0.067	0.648 ± 0.024
	LLaMa	K-Prec + Q2	0.712 ± 0.023	0.177 ± 0.028	0.189 ± 0.014	0.555 ± 0.038
	Mistral	K-Prec + Q2	0.691 ± 0.013	0.288 ± 0.015	0.290 ± 0.186	0.611 ± 0.015
	Mixtral	K-Prec + Q2	0.731 ± 0.032	0.276 ± 0.189	0.291 ± 0.178	0.672 ± 0.112
	ChatGPT	Spec. + K-Prec	0.711 ± 0.031	0.184 ± 0.012	0.182 ± 0.131	0.665 ± 0.021
	LLaMa	Spec. + K-Prec	0.704 ± 0.067	0.184 ± 0.034	0.187 ± 0.024	0.578 ± 0.024
	Mistral	Spec. + K-Prec	0.686 ± 0.021	0.281 ± 0.071	0.299 ± 0.045	0.591 ± 0.021
	Mixtral	Spec. + K-Prec	0.728 ± 0.021	0.282 ± 0.031	0.321 ± 0.141	0.711 ± 0.027
	ChatGPT	Spec. + Q2	0.721 ± 0.172	0.167 ± 0.021	0.174 ± 0.019	0.661 ± 0.027
	LLaMa	Spec. + Q2	0.718 ± 0.089	0.174 ± 0.056	0.188 ± 0.065	0.588 ± 0.041
	Mistral	Spec. + Q2	0.688 ± 0.129	0.267 ± 0.014	0.271 ± 0.011	0.578 ± 0.018
	Mixtral	Spec. + Q2	0.741 ± 0.121	0.282 ± 0.073	0.330 ± 0.021	0.676 ± 0.043
	ChatGPT	K-Prec + Q2 + Spec.	0.742 ± 0.014	0.221 ± 0.021	0.231 ± 0.178	0.684 ± 0.172
	LLaMa	K-Prec + Q2 + Spec.	0.740 ± 0.021	0.227 ± 0.018	0.238 ± 0.019	0.626 ± 0.016
	Mistral	K-Prec + Q2 + Spec.	0.762 ± 0.013	0.304 ± 0.016	0.326 ± 0.018	0.646 ± 0.221
	Mixtral	K-Prec + Q2 + Spec.	$\mathbf{0.781 \pm 0.012}$	$\mathbf{0.322 \pm 0.012}$	$\mathbf{0.344 \pm 0.017}$	$\mathbf{0.721 \pm 0.018}$

Table 6: Performance of ReMuSE with different combinations of rewards on the Meta-Reviewing Domain.

Domain	Model	Rewards	K-Prec	Q2-F1	Q2-NLI	Spec.
Product Buying	ChatGPT	-	0.497 ± 0.023	0.202 ± 0.014	0.203 ± 0.011	0.311 ± 0.024
	LLaMa	-	0.462 ± 0.022	0.203 ± 0.034	0.206 ± 0.044	0.448 ± 0.017
	Mistral	-	0.523 ± 0.015	0.225 ± 0.022	0.225 ± 0.019	0.427 ± 0.028
	Mixtral	-	0.461 ± 0.026	0.263 ± 0.039	0.323 ± 0.045	0.424 ± 0.076
	ChatGPT	K-Prec	0.539 ± 0.010	0.226 ± 0.012	0.240 ± 0.024	0.441 ± 0.012
	LLaMa	K-Prec	0.580 ± 0.032	0.236 ± 0.014	0.246 ± 0.321	0.541 ± 0.032
	Mistral	K-Prec	0.582 ± 0.036	0.267 ± 0.153	0.267 ± 0.242	0.475 ± 0.128
	Mixtral	K-Prec	0.622 ± 0.010	0.289 ± 0.011	0.287 ±0.152	0.529 ± 0.009
	ChatGPT	Q2 (F1 + NLI)	0.533 ± 0.029	0.287 ± 0.018	0.294 ± 0.015	0.438 ± 0.015
	LLaMa	Q2 (F1 + NLI)	0.544 ± 0.006	0.242 ± 0.018	0.257 ± 0.125	0.522 ± 0.025
	Mistral	Q2 (F1 + NLI)	0.593 ± 0.032	0.278 ± 0.036	0.282 ± 0.036	0.467 ± 0.042
	Mixtral	Q2 (F1 + NLI)	0.662 ± 0.004	0.335 ± 0.013	0.347 ± 0.011	0.525 ± 0.004
	ChatGPT	Spec.	0.643 ± 0.012	0.267 ± 0.019	0.276 ± 0.019	0.453 ± 0.039
	LLaMa	Spec.	0.621 ± 0.023	0.226 ± 0.027	0.239 ± 0.030	0.589 ± 0.029
	Mistral	Spec.	0.634 ± 0.037	0.258 ± 0.054	0.276 ± 0.069	0.602 ± 0.063
	Mixtral	Spec.	0.642 ± 0.015	0.342 ± 0.032	0.344 ± 0.035	0.592 ± 0.001
	ChatGPT	K-Prec + Q2	0.623 ± 0.017	0.243 ± 0.035	0.260 ± 0.121	0.463 ± 0.051
	LLaMa	K-Prec + Q2	0.631 ± 0.046	0.234 ± 0.129	0.237 ± 0.038	0.552 ± 0.042
	Mistral	K-Prec + Q2	0.641 ± 0.043	0.242 ± 0.117	0.264 ± 0.015	0.588 ± 0.052
	Mixtral	K-Prec + Q2	0.671 ± 0.432	0.326 ± 0.012	0.341 ± 0.028	0.602 ± 0.016
	ChatGPT	Spec. + K-Prec	0.612 ± 0.015	0.241 ± 0.016	0.311 ± 0.178	0.524 ± 0.028
	LLaMa	Spec. + K-Prec	0.582 ± 0.016	0.248 ± 0.018	0.254 ± 0.048	0.581 ± 0.063
	Mistral	Spec. + K-Prec	0.633 ± 0.061	0.241 ± 0.017	0.263 ± 0.018	0.572 ± 0.061
	Mixtral	Spec. + K-Prec	0.664 ± 0.038	0.327 ± 0.019	0.341 ± 0.026	0.585 ± 0.042
	ChatGPT	Spec. + Q2	0.652 ± 0.041	0.267 ± 0.045	0.287 ± 0.078	0.535 ± 0.056
	LLaMa	Spec. + Q2	0.621 ± 0.064	0.257 ± 0.054	0.268 ± 0.068	0.591 ± 0.017
	Mistral	Spec. + Q2	0.641 ± 0.074	0.256 ± 0.076	0.322 ± 0.059	0.622 ± 0.078
	Mixtral	Spec. + Q2	0.676 ± 0.051	0.358 ± 0.067	0.361 ± 0.071	0.621 ± 0.051
	ChatGPT	K-Prec + Q2 + Spec.	0.671 ± 0.014	0.284 ± 0.151	0.314 ± 0.115	0.555 ± 0.157
	LLaMa	K-Prec + Q2 + Spec.	0.637 ± 0.015	0.286 ± 0.131	0.299 ± 0.141	0.581 ± 0.231
	Mistral	K-Prec + Q2 + Spec.	0.697 ± 0.051	0.339 ± 0.182	0.351 ± 0.132	0.604 ± 0.214
	Mixtral	K-Prec + Q2 + Spec.	**0.713 ± 0.141**	**0.372 ± 0.121**	**0.404 ± 0.151**	**0.652 ± 0.191**

Table 7: Performance of ReMuSE on the Product Buying Domain. Best results are indicated in **bold**.

Domain	Model	Rewards	K-Prec	Q2-F1	Q2-NLI	Spec.
Debates	ChatGPT	-	0.374 ± 0.011	0.171 ± 0.121	0.202 ± 0.141	0.398 ± 0.141
	LLaMa	-	0.377 ± 0.021	0.177 ± 0.214	0.209 ± 0.151	0.427 ± 0.115
	Mistral	-	0.406 ± 0.022	0.223 ± 0.113	0.246 ± 0.133	0.443 ± 0.122
	Mixtral	-	0.411 ± 0.015	0.208 ± 0.141	0.226 ± 0.115	0.451 ± 0.114
	ChatGPT	K-Prec	0.444 ± 0.006	0.180 ± 0.131	0.215 ± 0.151	0.482 ± 0.221
	LLaMa	K-Prec	0.414 ± 0.012	0.187 ± 0.152	0.214 ± 0.042	0.529 ± 0.033
	Mistral	K-Prec	0.486 ± 0.004	0.235 ± 0.151	0.244 ± 0.041	0.547 ± 0.003
	Mixtral	K-Prec	0.483 ± 0.009	0.246 ± 0.015	0.241 ± 0.141	0.581 ± 0.033
	ChatGPT	Q2 (F1 + NLI)	0.421 ± 0.028	0.194 ± 0.020	0.228 ± 0.024	0.481 ± 0.027
	LLaMa	Q2 (F1 + NLI)	0.412 ± 0.027	0.216 ± 0.039	0.224 ± 0.045	0.518 ± 0.033
	Mistral	Q2 (F1 + NLI)	0.487 ± 0.039	0.251 ± 0.043	0.277 ± 0.045	0.552 ± 0.015
	Mixtral	Q2 (F1 + NLI)	0.504 ± 0.009	0.268 ± 0.034	0.288 ± 0.010	0.574 ± 0.022
	ChatGPT	Spec.	0.414 ± 0.017	0.177 ± 0.018	0.186 ± 0.027	0.494 ± 0.018
	LLaMa	Spec.	0.400 ± 0.031	0.181 ± 0.047	0.190 ± 0.044	0.540 ± 0.032
	Mistral	Spec.	0.484 ± 0.024	0.249 ± 0.046	0.250 ± 0.048	0.581 ± 0.018
	Mixtral	Spec.	0.491 ± 0.019	0.252 ± 0.012	0.261 ± 0.017	0.601 ± 0.010
	ChatGPT	K-Prec + Q2	0.419 ± 0.057	0.208 ± 0.054	0.214 ± 0.043	0.526 ± 0.051
	LLaMa	K-Prec + Q2	0.411 ± 0.064	0.192 ± 0.051	0.201 ± 0.054	0.537 ± 0.018
	Mistral	K-Prec + Q2	0.473 ± 0.051	0.262 ± 0.043	0.278 ± 0.052	0.574 ± 0.087
	Mixtral	K-Prec + Q2	0.494 ± 0.064	0.268 ± 0.018	0.267 ± 0.064	0.628 ± 0.065
	ChatGPT	Spec. + K-Prec	0.422 ± 0.052	0.201 ± 0.044	0.209 ± 0.116	0.484 ± 0.054
	LLaMa	Spec. + K-Prec	0.428 ± 0.089	0.194 ± 0.062	0.205 ± 0.067	0.542 ± 0.144
	Mistral	Spec. + K-Prec	0.489 ± 0.057	0.244 ± 0.067	0.259 ± 0.089	0.584 ± 0.056
	Mixtral	Spec. + K-Prec	0.484 ± 0.078	0.252 ± 0.056	0.252 ± 0.018	0.624 ± 0.076
	ChatGPT	Spec. + Q2	0.448 ± 0.197	0.201 ± 0.069	0.212 ± 0.068	0.502 ± 0.074
	LLaMa	Spec. + Q2	0.421 ± 0.079	0.221 ± 0.168	0.248 ± 0.156	0.551 ± 0.018
	Mistral	Spec. + Q2	0.486 ± 0.058	0.248 ± 0.188	0.268 ± 0.076	0.591 ± 0.054
	Mixtral	Spec. + Q2	0.510 ± 0.098	0.278 ± 0.088	0.281 ± 0.045	0.621 ± 0.077
	ChatGPT	K-Prec + Q2 + Spec.	0.501 ± 0.016	0.221 ± 0.114	0.232 ± 0.168	0.535 ± 0.132
	LLaMa	K-Prec + Q2 + Spec.	0.441 ± 0.131	0.234 ± 0.178	0.244 ± 0.132	0.565 ± 0.151
	Mistral	K-Prec + Q2 + Spec.	0.501 ± 0.112	0.274 ± 0.213	0.291 ± 0.121	0.627 ± 0.189
	Mixtral	K-Prec + Q2 + Spec.	$\mathbf{0.522 \pm 0.131}$	$\mathbf{0.277 \pm 0.122}$	$\mathbf{0.298 \pm 0.152}$	$\mathbf{0.628 \pm 0.121}$

Table 8: Performance of ReMuSE on the Debates Domain. Best results are indicated in **bold**.

Prompt for Refinement (Debates)

Instruction: Given the feedback, knowledge source, and dialogue improve the dialogue based on the feedback provided. The output should just be the new dialogue.

Knowledge Source: ***For arguments****: 'Let's talk about what would happen to real people in real families if the U.S. were to go down this untested and quite radical path of eliminating the minimum wage as our opponents urge...', **Against Arguments**: 'We need to abolish the minimum wage. And there's only one argument that matters on this issue, and it's the moral argument. Does the minimum wage make the world a better place?...'*

Feedback: *It could benefit from more specific and concise answers to the questions asked by the decision-maker. 1. The Q2 F1 score could be improved by providing more specific examples and evidence to support the arguments presented. For example, when discussing the argument for abolishing the minimum wage, the dialogue agent could provide examples of how higher labor costs have negatively impacted businesses in the past. 2. The Q2 NLI score could be improved by ensuring that the dialogue agent's responses are coherent and make sense in the context of the conversation. For example, when discussing the argument for keeping the minimum wage, the dialogue agent could provide more specific examples of how a living wage would benefit low-income families.*

Dialogue: *Decision Maker: Hello, can you tell the arguments for and against debating for minimum wages? \n Dialogue Agent: The debate over abolishing the minimum wage centers on two perspectives. Supporters claim it harms businesses and argue for market-driven wages without government interference. Opponents assert the minimum wage is crucial to shield low-income workers from exploitation, contending that without it, wages might be insufficient for basic needs. \n Decision Maker: ..*

A.7 Prompts for Generic and Actionable feedback

We discuss the prompts used in our work here:

Prompt for Generic feedback (Meta-Reviewing)

Instruction: Given the knowledge source and the dialogue, please provide **feedback** to improve the dialogues. The feedback should just be for the overall dialogue and should start with 'Feedback:'.

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ..., **Review 2**: .., **Review 3**: ...*

Dialogue: *Meta-Reviewer: Hello Dialogue Agent. Can you tell me more about this paper? \n Dialogue Agent: Ofcourse! This is a paper about CodexGLUE, a machine learning benchmark dataset for code understanding and generation, \n MetaReviewer: ...*

Prompt for Actionable feedback (Meta-Reviewing)

Instruction: Given the knowledge source and the dialogue, please provide **actionable feedback** to improve the dialogues. The feedback should be for the overall dialogue and start with 'Feedback:.' The feedback should improve the groundedness of the dialogue agent's utterances to the knowledge source. The feedback should also increase the specificity (higher technical details) of utterances for the dialogue agent and the meta-reviewer.

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ..., **Review 2**: .., **Review 3**: ...*

Dialogue: *Meta-Reviewer: Hello Dialogue Agent. Can you tell me more about this paper? \n Dialogue Agent: Ofcourse! This is a paper about CodexGLUE, a machine learning benchmark dataset for code understanding and generation, \n MetaReviewer: ...*

A.8 Cross-Domain analysis for the performance of ReMuSE

Table 6 shows the detailed performance of ReMuSE with multiple combinations of rewards for the meta-reviewing domain. We discuss the cross-domain analysis of ReMuSE on two other decision-making domains namely debates and product reviews below.

A.8.1 Domains

In order to test the robustness of our approach, ReMuSE, we experiment with two other domains of decision-making based on lengthy input documents. In *product-buying*, a buyer needs to decide for or against a product based on the provided customer reviews. In *debates*, a listener must decide on one of the stances based on the provided arguments.

In this section, we initially discuss the datasets we employed to perform the analysis, followed by a discussion on our experimental findings.

A.8.2 Data

For *product-buying*, we use the Helpful Reviews Dataset (Gamzu et al., 2021). The reviews are available for 123 products, and each review comes annotated with a helpfulness score. Following the recommendation of Gamzu et al. (2021), we consider all the reviews scored above 1, which are the most helpful for making buying decisions. For *debates*, we use the IQ2 Debates Dataset (Zhang et al., 2016). The dataset consists of Oxford-Style Debates on 108 topics. Each debate is a dialogue between multiple parties namely, moderator, pro-side, and con-side. For ease of use, we only con-

sider the opening arguments from the pro- and con-sides of the debate for each topic.

A.8.3 Results

In the case of product buying, we use all the helpful review sentences available for a product as the knowledge source, whereas the arguments from pro and con-sides of a topic in case of debates. We use similar prompts as in Sec §3.2 but simulate different scenarios for both domains and then apply our method as is. The detailed prompts for initial dialogue generation, feedback, and refinement for both the domains are provided in §A.6.1, §A.6.2 and §A.6.3 respectively. We report the results for the *product-buying* and *debates* domain in Tables 7 and 8 in §A.8 respectively. The best results are attained by using an amalgamation of all the refinement metrics, similar to the meta-reviewing domain. We observe a similar trend in terms of the best-performing model as well where the Mixtral model excels across all the domains.

A.9 Ablation Study

A.9.1 Single vs Joint Reward Optimization

We show the scores for single vs joint score optimization in Table 9. We observe massive improvements when using the amalgamation of all the refinement scores.

Reward	K-Prec	Q2-F1	Q2-NL	Spec.
K-Prec	$0.70_{.01}$	$0.27_{.02}$	$0.26_{.02}$	$0.65_{.03}$
Q2	$0.72_{.01}$	$0.27_{.02}$	$0.27_{.01}$	$0.70_{.03}$
Spec.	$0.70_{.01}$	$0.29_{.02}$	$0.27_{.01}$	$0.69_{.02}$
K-Prec + Q2	$0.73_{.03}$	$0.28_{.15}$	$0.29_{.18}$	$0.67_{.11}$
Spec. + K-Prec	$0.73_{.02}$	$0.28_{.03}$	$0.32_{.14}$	$0.71_{.03}$
Spec. + Q2	$0.74_{.12}$	$0.28_{.07}$	$0.33_{.02}$	$0.68_{.04}$
ReMuSE	$\mathbf{0.78}_{.01}$	$\mathbf{0.32}_{.01}$	$\mathbf{0.34}_{.02}$	$\mathbf{0.72}_{.02}$

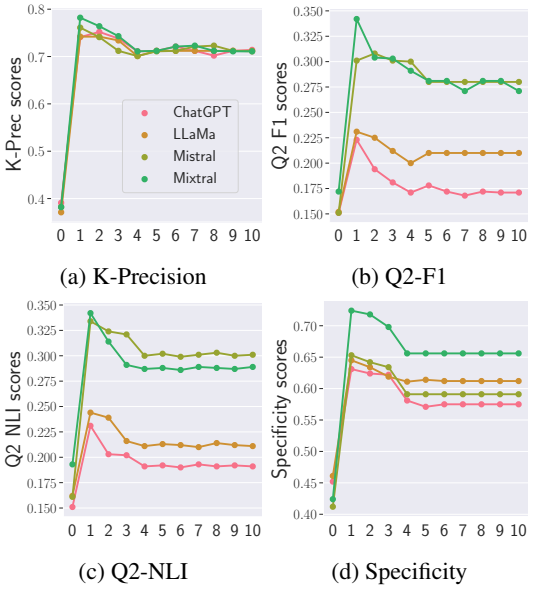
Table 9: Single vs Joint-Score Optimization of ReMuSE for our best-performing model, **Mixtral** in terms of various refinement metrics.

A.9.2 Multi-Iterative Refinement

In Fig 5, we plot the scores for each iteration of ReMuSE up to 10 epochs and observe that the scores plateau after the first epoch, indicating superior feedback quality in the first round, consistent with prior work (Madaan et al., 2023).

A.9.3 Generic vs Actionable vs Multi-Reward FeedBack

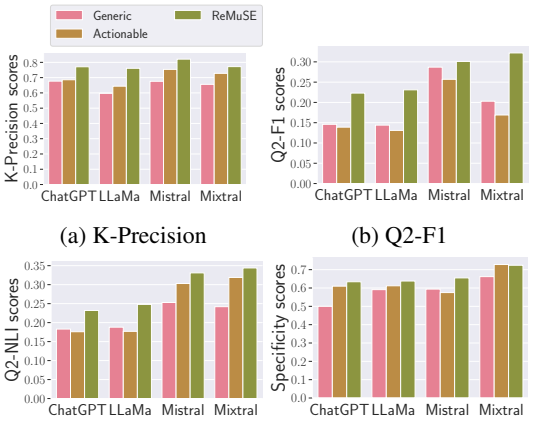
We plot the various metrics for these multiple feedback strategies in Fig 6. We observe that the feedback-prompting in ReMuSE leads to the maximum improvements. We also show a few examples



(a) K-Precision (b) Q2-F1

(c) Q2-NLI (d) Specificity

Figure 5: For the iterative refinement version of ReMuSE, we plot the (a) K-Precision, (b) Q2-F1, (c) Q2-NLI, and, (d) Specificity scores with varying iterations $(0-4)$. All the results are averaged over 3 prompts.



(a) K-Precision (b) Q2-F1

(c) Q2-NLI (d) Specificity

Figure 6: Performances for the different feedback types Generic, Actionable, and ReMuSE. We plot the (a) K-Precision, (b) Q2-F1, (c) Q2-NLI, and, (d) Specificity scores for each (averaged over 3 prompts).

of the feedback generated for a dialogue using the prompts discussed above and our proposed method, ReMuSE in Table 10.

A.10 Performance of Dialogue Agents on Meta-Reviewing

We report the performance of all the models for the dialogue agent response generation task in Table 11. We observe that Flan-T5 outperforms all the models across the board.

Strategy	Feedback
Generic	The agent mostly summarized the concerns raised by the reviewers without suggesting any specific actions that could be taken to address those concerns. In the future, the agent could provide more specific suggestions or questions to help the meta-reviewer make a more informed decision. Additionally, the agent could try to paraphrase or restate the concerns raised by the reviewers in a clearer or more concise way to facilitate the conversation.
Actionable	Overall, it is advisable for the authors to further clarify their proposed algorithm by addressing the issues raised by the reviewers. For instance, they can provide a better definition of transportation polytope in Eq. (2) and explain the process of generating two views of the input images in more detail. It is also recommended that they compare their method with the state-of-the-art consensus clustering methods that are not deep learning-based as suggested by Reviewer 1. To address the concerns over the quality of the experiments, the authors can provide a more thorough experimental investigation and comparison, as well as a discussion of the novelty and contributions of their method with respect to similar methods proposed in prior work.
ReMuSE	Here are some suggestions to improve the dialogue: 1. Provide a more detailed summary of the reviewers' concerns, such as the specific areas where the paper is lacking in novelty or clarity. For example, the dialogue could mention the reviewers' comments about the outdated literature, vague algorithm description, and lack of analysis of results. 2. Offer specific recommendations to address the reviewers' concerns. For instance, the dialogue could suggest that the authors update the literature review with more recent work on consensus clustering, provide a more detailed explanation of the algorithm, and perform additional analysis of the results. 3. Highlight the positive aspects of the paper, such as the novelty of combining ensemble methods with deep clustering. The dialogue could also mention the reviewers' comments about the potential effectiveness of the proposed ensemble deep clustering algorithm.

Table 10: Different types of feedback generated for the same dialogue using multiple feedback prompting strategies.

Models	$(y, \hat{y})$		$(y, \mathcal{K})$		
	BL.	**BT Sc.**	**K-Prec**	**Q2-F1**	**Q2-NL**
OPT	15.4	54.5	57.6	36.2	40.2
T5	18.4	58.1	63.4	39.1	45.4
Flan-T5	**20.4**	**60.3**	**68.2**	**42.2**	**47.3**
ChatGPT	12.2	45.2	42.1	25.1	25.4
ReMuSE + K-Prec	13.5	47.0	65.8	40.4	46.2
ReMuSE + Q2	13.7	47.2	65.1	41.1	46.6
ReMuSE + Spec.	13.9	47.5	66.9	41.3	45.9
ReMuSE + Spec. + K-Prec	14.2	48.1	66.0	41.5	46.5
ReMuSE + K-Prec + Q2	14.3	48.4	67.2	42.0	46.8
ReMuSE + Q2 + Spec	14.1	48.2	67.4	42.1	46.9
ReMuSE + all	14.5	49.0	67.6	42.2	47.1

Table 11: Performance of models in the dialogue agent response generation task. BLEU (BL.) and BERT Score (BT Sc.) measure the similarity of the response with the ground truth. K-Prec and Q2 scores evaluate the factual consistency of the response with the Knowledge Source.

A.10.1 Correlation of human evaluation metrics

We plot the correlation of the different metrics used for evaluating the dialogue agent's responses, namely, fluency, faithfulness, relevance, objectivity, and helpfulness, in Fig 8. We observe a strong correlation between faithfulness, objectivity, and helpfulness. This further indicates a low tolerance for unfaithful responses in the decision-making process. The models finetuned on our data generate more faithful responses as demonstrated through the scores in human evaluation in Table 4 in Sec §4.2.

A.11 Diversity of decisions in Dagstuhl dataset

The Dagstuhl Dataset consists of 4 papers meta-reviewed by 4 senior academicians. The paper acceptance decisions is displayed in Figure 4. Across the 4 papers, only 1 paper had a unanimous reject decision, while all the other papers received diverse verdicts.

A.12 Resulting Dataset

As one of the contributions of this work, we make the synthetically generated dataset for the best-performing ReMuSE configuration publicly available. We release $3,064$ meta-reviewing, 128 product-buying, and 108 debate dialogues, respectively. We display the basic statistics for our resulting dataset in Table 12. Meta-Reviewing dialogues show the most diversity (unigram vocabulary size) in agent and meta-reviewer utterances, reflecting the specific terminology associated with the peer-reviewing domain. The debates domain closely follows due to diverse topics, while the product-buying domain exhibits the least diversity, reflecting the domain's simplicity. Meta-Reviewing dialogues are notably longer, with more turns depicting the longer reasoning steps in expert tasks. The turn distributions in our dataset are highlighted in Figures 10, 11, 12 for meta-reviewing, product buying, and debates datasets, respectively.

A.13 Hyperparameters for Dialogue Response Generation

All models are fine-tuned for 10 epochs. We perform a grid search on learning rates $\{1e-4, 3e-4, 5e-5, 6e-5\}$ and select the best model based on the validation set performance. The results are averaged over 3 random seeds. For ReMuSE, we use the best-performing model, Mixtral, and update the responses of ReMuSE for 1 iteration, given our prior findings. We adapt the same prompts as in A.6 for the response generation task.

Response generation Prompt (Meta-Reviewing)

Instruction: Generate a response corresponding to a dialogue for a conversation between a meta-reviewer and a dialogue agent for reviews about a paper. The The dialogue agent has access to the dialogue history, reviews, and the type of paper. The dialogue agent should just answer from the reviews. A dialogue agent should never recommend anything/ give any opinions/ decide anything for the paper. A dialogue agent has no conscience of its own.

Title: [title]

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: .., **Review 3**: ...* **Dialogue History:** ***Meta-reviewer****: Hello, dialogue agent. Thank you for joining me today to discuss the reviews for the paper [title]. Can you please start by summarizing the main contributions of the paper and how it addresses the problem it sets out to solve?*

Prompt for Response generation ReMuSE feedback

Instruction: Given the knowledge source, dialogue history and the response, please provide actionable feedback to improve the response. The feedback should be for the response and should start with 'Feedback:'. A Q2 F1 score, Q2 NLI score, KPrecision, and specificity scores. The Q2 and Kprecision scores measure the groundedness of the response to the provided knowledge source. The Specificity scores measure the technicality of the response. The feedback should try to improve all scores for the dialogue agent

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: .., **Review 3**: ...*

Dialogue History: ***Meta-reviewer****: Hello, dialogue agent. Thank you for joining me today to discuss the reviews for the paper [title]. Can you please start by summarizing the main contributions of the paper and how it addresses the problem it sets out to solve?*

Response: ***Dialogue Agent****: Sure. The paper proposes a method based on importance and coherence to evaluate whether a partitioning of a neural network exhibits modular characteristics. F1:0.12, NLI:0.14 and Specificity: 0.1*

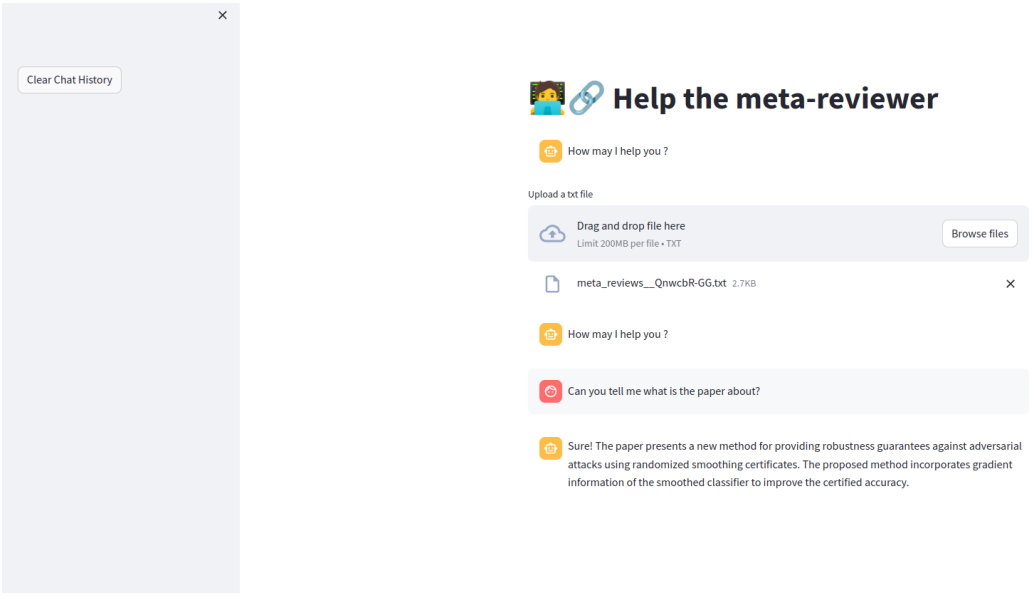


Figure 7: Screenshot of the Interface for assisting meta-reviewers with a dialogue agent

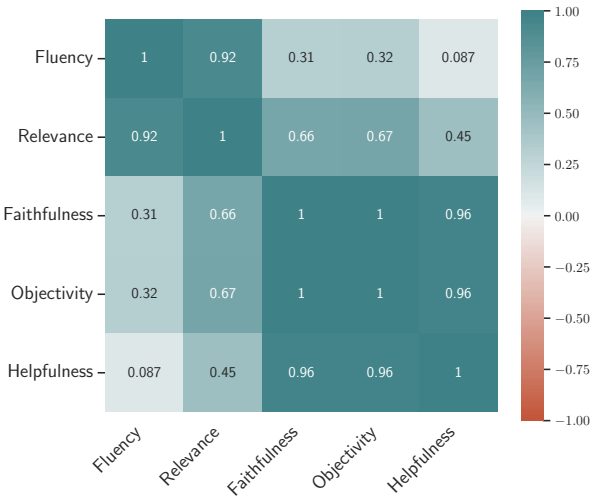


Figure 8: Correlation of the human evaluation metrics. We observe the strongest correlation between Helpfulness, Faithfulness, and Objectivity.

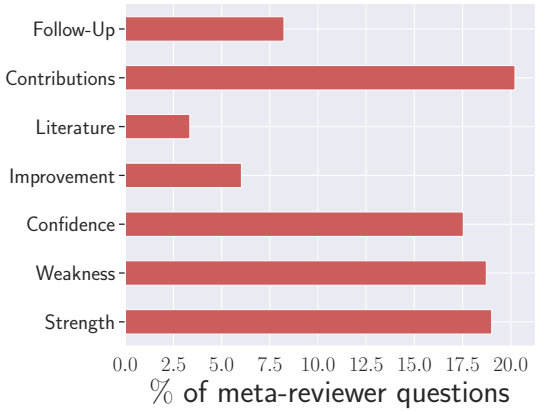


Figure 9: Distribution of meta-reviewer questions in our dataset.

Prompt for Response generation Re-MuSE Refinement

Instruction: Given the feedback, knowledge source, dialogue history and the previous response, improve the response based on the feedback provided. The output should just be the new dialogue.

Knowledge Source: ***Review 1****: The paper proposes a benchmark to measure the quality of the systems in the growing domain of code intelligence ...,* ***Review 2****: .,* ***Review 3****: ...*

Feedback: *It would be helpful to provide a clearer explanation of the tasks' relevance to software development in response to reviewer one's concerns. Additionally, it may be useful to mention that the paper discusses the effectiveness of the baselines and provides an ablation study. Finally, it would be good to acknowledge reviewer three's comment explicitly and mention any plans to revise the descriptions of the datasets.*

Dialogue history: ***Meta-reviewer****: Hello, dialogue agent. Thank you for joining me today to discuss the reviews for the paper [title]. Can you please start by summarizing the main contributions of the paper and how it addresses the problem it sets out to solve?*
Previous Response: ***Dialogue Agent****: Sure. The paper proposes a method based on importance and coherence to evaluate whether a partitioning of a neural network exhibits modular characteristics.*

A.14 Meta-Reviewer Question Distribution

To quantify the different types of questions, we quantitatively analyze the different meta-reviewer questions in our dataset. To come up with the categories, we use the 'Dagstuhl' dataset and annotate the questions in that dataset. We find that most of the meta-reviewer questions are related to Contributions, Strengths, Weaknesses, Confidence, Avenues for Improvement, and Literature Survey and Follow-Up. We then prompt GPT 3.5-turbo to provide us with these labels for all the questions in our dataset. The distribution of questions in our dataset is provided in Fig 9. We observed that questions related to contributions, strengths, weaknesses, contributions, and confidence ($\sim$ 20%) occur across all dialogues. These questions are the most impor-

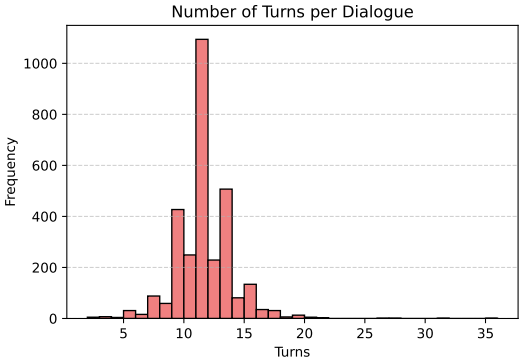


Figure 10: Distribution of Number of turns in Meta-Reviewing Dataset

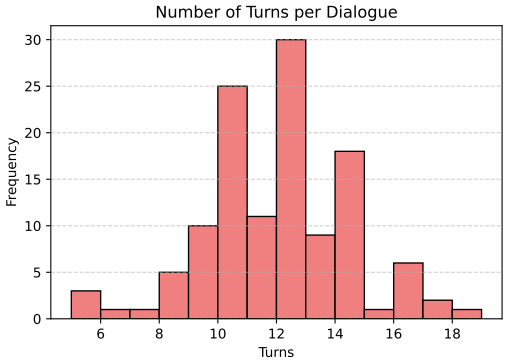


Figure 11: Distribution of Number of turns in Product Buying Dataset

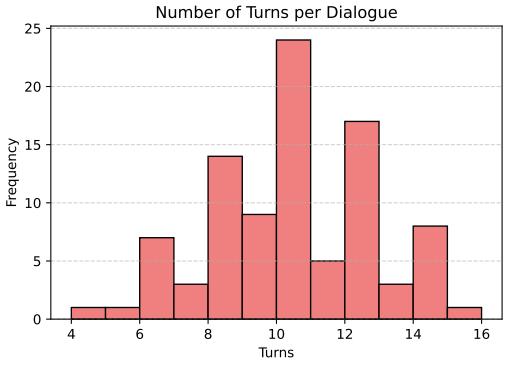


Figure 12: Distribution of Number of turns in Debates Dataset

tant aspects for consideration while writing a meta-review, which is in line with the findings of Santu et al. (2024). Thus, the questions in our dataset can serve as an effective resource for serving as a proxy meta-reviewer while training dialogue agents.

A.15 Error Analysis of the dialogue agent responses

Following Li et al. (2023b), we quantitatively analyze a set of 50 dialogues for error types in the dialogue agent's responses. Out of 150 dialogue agent turns, we observed 50 responses that may be considered erroneous. We identified four different kinds of errors in the responses: 1) **Irrelevant answers**: 75% errors are incurred when the dialogue agent tries to answer questions that are based on external knowledge. For instance, for a meta-reviewer question on the literature survey of the paper, the dialogue agent answers about the novelty of the method in relation to the prior work as available in the reviewer comments. However, these responses are short, and the dialogue agent evades the answer by providing a partial/unsatisfactory answer; 2) **Inappropriate answers**: 10% of the errors are incurred because of the lack of knowledge in using comparative adjectives. For instance, responses to questions regarding the confidence of a particular reviewer are often answered with the confidence of all the reviewers. This question is particularly asked as a follow-up to the severity of the weakness stated by the reviewers. However, the dialogue agent answers with the confidence of other reviewers using adjectives such as 'low' or 'high'. These responses are sometimes confusing since confidence levels, such as 4 out of 5, are not particularly low but are conceived as low because of the higher confidence of other reviewers. 3) **Repitition**: 10% of responses have repeated facts as other answers and lack diversity. 4) **Lack of Neutrality**: In 5% of the cases, the dialogue agent agrees with the meta-reviewer's evaluation of the paper, which is not intended since the dialogue agent should state its neutral stance. However, we did not observe any cases of disagreement with the meta-reviewer, which justifies the refinement metrics used in the dialogue generation.

As a part of our research design, we induce faithfulness and technical specificity in the LLM generations. Thus, errors related to irrelevant answers occur due to increased faithfulness in the generations and the necessity to stick to the source content. The other errors are mainly related to limitations

Domain	# dial.	# DA. tok. (av.)	# Seek. tok. (av.)	# tur.
Meta-Rev.	**3,064**	**54.66**	**30.95**	**12.24**
Product-Buy.	123	28.87	13.49	11.55
Debates	108	48.51	18.16	10.0

Table 12: Basic Statistics for the generated dataset. The table shows the number of dialogues, the average number of tokens for the dialogue agent and seeker utterances, and the number of turns.

in LLMs in using comparative adjectives and the inherent pre-training-based alignment to agree with humans. However, none of the errors contradict our research design, and all the responses are still technical and faithful. This justifies the use of the dataset to train dialogue agents with higher faithfulness and technicality.

A.16 Human Evaluation Instruction for evaluating full dialogues

Please evaluate the dialogue in Table 14 with respect to the reviews in Sec §A.22.1 on a Likert Scale of $1-4$ over the following dimensions:
Co-operativeness means that the response of the dialogue agent is coherent with the previous turn and does not try to mislead the interlocutor or act unhelpfully.
Coherence how coherent and consistent is the overall dialogue.
Engagingness implies the dialogue agent engages the interlocutor by prompting further replies and helps move the conversation forward.
Plausibility evaluate the groundedness of the dialogue agent's response. The measure implies how reasonable the information is about the reviews and the paper as provided by the assistant.

A.17 Human Evaluation Instruction for evaluating dialogue responses

Please evaluate the response, $u_t + 1$ in connection to the Knowledge source, $\mathcal{K}$ and dialogue history $[u_1, u_2, \ldots u_t]$ on a Likert scale of $1-5$ over the following dimensions:
Fluency How fluent is the language usage in the response?
Relevance How appropriate is the response to the dialogue history?
Faithfulness Is the response grounded/faithful to the knowledge source, $\mathcal{K}$.
Objectivity Is the answer factually correct rather than providing personal opinions and feelings?
Helpfulness How useful is this response if you had to decide the acceptance or rejection of this paper based on this response?

Knowledge Source, $\mathcal{K}$

Review 1: This paper proposes a new type of generative models with a new inference method of latent variables. ..., *Review 2*: The paper proposes GONs which seek to build a generative model with an "implicit" encoder that comes essentially for free with the use of a few re-parameterization tricks .., *Review 3*: This paper introduces a "new" inference method for autoencoder-type models, where the encoder is taken as a gradient of the decoder with respect to a zero-initialized latent variable ...

Dialogue History, $[u_1, u_2, \ldots u_t]$
Meta Reviewer: Hello, can you summarize the main contributions of this paper "Gradient Origin Networks"?
Dialogue Agent: Yes, of course. This paper proposes a new type of generative models called Gradient Origin Networks which use a new inference method of latent variables. Specifically, the gradient of latent variables with respect to the zero vector is taken as the inferred latent variables.
Meta Reviewer: I see. Can you tell me more about the pros and cons mentioned in the reviews?]

Response, $u_t + 1$

The reviews describe that the proposed method is easy and straightforward to implement. However, the model assumption that the one step gradient from zero vector equals to latent vector is quite limited and greatly constrains the model expressiveness. Also, the formulation needs to be carefully checked. The experimental results are not sufficient to demonstrate the efficacy. Need more quantitative analysis and experiments on more challenging datasets. The claim that it saves parameters compared to VAE is also confusing.

A.18 Annotator Details

Out of the 10 expert annotators in our study, 4 annotators are 2nd-year Ph.D students while 2 annotators are 1st-year Ph.D students, and the other 4 annotators are 3rd-year Ph.D students. 5 annotators are from India, 4 annotators are from Germany, and 1 annotator is from Belgium. All of them are fluent English speakers who have experience in NLP peer-reviewing. The non-expert annotators are from the fields of Mathematics, Railway Engineering, Archi-

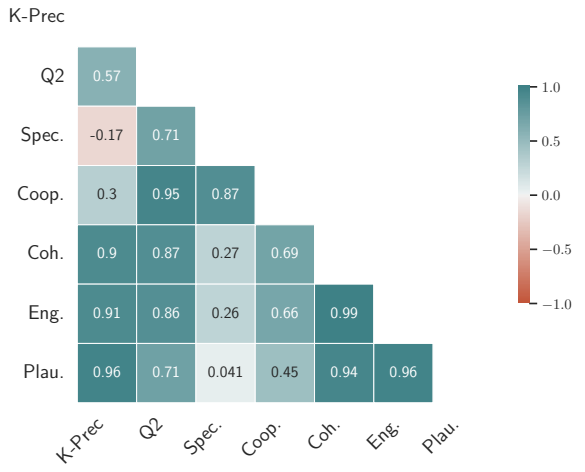


Figure 13: Correlation between human and automated evaluation metrics

tecture, Civil Engineering, and Medieval History. Three of the annotators are from Germany, one from Hong Kong, and the other from Italy.

A.19 Correlation between human and automated metrics

Since human evaluation metrics are widely used and easily interpretable (Macina et al., 2023; Rashkin et al., 2021), we analyze their correlation with automated metrics. We compute correlations between human and automated evaluations (Tables 1 and 2), with results shown in Fig 13. We find that automated metrics closely aligned with human metrics exhibit strong correlations (e.g., K-Prec and Q2: 0.95), while those measuring different aspects, such as Specificity and Plausibility, show lower correlation (0.04). Lower correlations within the same category further highlight the diversity of captured aspects.

A.20 Disagreement analysis for Human Evaluation in response generation

We analyzed the instances where the scores varied by 2 or more points (e.g., one annotator scores a response 1 for fluency, while another scores it 4). Out of the 240 responses (60 from each model), the most significant disagreements occurred when rating ChatGPT's responses (12 out of 60). Of these 12 responses, 7 are related to areas of improvement, 4 to the literature review, and 1 to a weakness. All of these responses are notably wordy, averaging 100 words, and often diverge off-topic, a pattern also observed in previous studies (Lozić and Štular, 2023). While these responses are fluent, the

Metrics	w/o Dial. Agent	w. Dial. Agent
Expert		
Content Relevance	3.42	**3.87**
Decision Correctness	**0.70**	0.62
Coverage	3.21	**3.76**
Non-expert		
Content Relevance	3.10	**3.65**
Decision Correctness	0.58	**0.60**
Coverage	2.95	**3.50**

Table 13: Human evaluation of meta-reviews written by experts and non-experts, with and without dialogue agent assistance. Content Relevance and Coverage are rated on a scale of 1–4. Decision Correctness is binary (0/1).

variation in objectivity scores between annotators suggests differing perspectives on the content. In contrast, we observed much higher agreement on responses from other models, excluding ChatGPT, with scores for relevance (0.75), faithfulness (0.78), objectivity (0.78), fluency (0.72), and helpfulness (0.70). For ChatGPT, the agreement scores were lower: fluency (0.52), relevance (0.65), faithfulness (0.68), objectivity (0.68), and helpfulness (0.68). Since these measures are used for evaluation rather than to generate gold standards, calibration is performed using averaging, as done in several prior works (Dziri et al., 2022; Rashkin et al., 2021).

A.21 Interface for Conversation and Dialogue Agent Evaluation

We developed the interface using Streamlit to interact with the dialogue agent as shown in Fig 7.[23] In Table 13, we observe higher scores for content relevance and Coverage while using a dialogue agent for both expert and non-expert meta-reviewers. However, decision correctness was lower while using a dialogue agent for the experts, though they perfectly aligned with the non-experts.

[23] https://streamlit.io/

A.22 Example Dialogues from our Dataset

We show a few reviews from our dataset along with the dialogue generated using ReMuSE below.

A.22.1 Meta-Reviewing

The title of the paper is not revealed to retain anonymization. The type of the paper is “long".

Review 1: Paper is on modeling the prediction of ancestor relation between names of science institutions. This is on the GRID dataset which already has some hierarchical information. The proposed approach is set-based models (with neural encodings) where the overlap between two names is measured by set overlap at the unigram level. In extended experiments additional metadata like address and type of institution are also incorporated into the model (which contribute a lot to the improvements). A set of simple to intermediate baseline along with different thresholds of token overlap has been tested and the proposed model shows strong improvement in the MAP metric. Paper has a decent writing and structure. Problem and the approach has been explained and motivated well with descriptive examples. [...]

Review 2: The paper shows how to infer the organisational structure of an institution. That is, it presents a model for predicting the is-ancestor relationships of institutions based on their string names. To this end, it makes use of Set-Transformers to model the token overlap between the institution names. This use is nice but also not highly original. The experimental evaluation is on a single dataset only. While the authors do present some examples, and overall hierarchy or something that provides some more insights into the learned model should be provided in order to show potential issues with transitivity and connected components. The evaluation only considers known pairs. But an organisational structure should also be consistent. That is, the interesting motivation provided in the intro is not met in the experimental evaluation. Furthermore, the experimental protocol is unclear [...]

Review 3: The paper addresses the prediction of the hierarchical structure of organizations / institutes. The authors develop a new dataset, automatically derived from GRID (global research identifier database), and compare a set-based model against a few baseline approaches. While the task is well-defined and the dataset could potentially be interesting for the community, I have a few doubts regarding the experimental setup (to be more specific: on the choice of baseline models, on the evaluation on the test set and on the final results). Pro: The task sounds interesting and challenging. It could encourage researchers to build and enhance models that combine knowledge from different sources. Con: The task is presented as a knowledge base completion task. [...]

Meta-reviewing Dialogue	
Meta-Reviewer	Hello, I'm reviewing a paper on predicting institution hierarchies with set-based models. Can you tell me a little bit about the paper?
Dialogue Agent	Sure! The paper presents a new approach to predicting the hierarchical structure of institutions using set-based models with neural encodings. The approach is based on the GRID dataset, which already has some hierarchical information.
Meta-Reviewer	That sounds interesting. Can you tell me more about the dataset and the experimental setup?
Dialogue Agent	Sure! The dataset is the GRID dataset, which is a global research identifier database. The authors have constructed a new dataset by pairing institution names with their corresponding hierarchy information. They have also compared their approach against a few baseline approaches and presented some examples.
Meta-Reviewer	Okay, that sounds good. But I have a few doubts regarding the experimental setup. For example, I would have expected a comparison with well-known baseline models from the knowledge base completion literature, such as TransE. And I'm not sure if the manually extracted test set is complete. Can you tell me more about this?
Dialogue Agent	Yes, the authors have mentioned some of these concerns in their paper. They have developed a new dataset, and the task is well-defined. However, they have not compared their approach with well-known baseline models from the knowledge base completion literature, such as TransE. And the manually extracted test set may not be complete, which may affect the evaluation.
Meta-Reviewer	I see. That's a good point. What about the token-based models? Are they a promising choice for the task?
Dialogue Agent	The authors have presented some examples in Table 3 to show the effectiveness of their approach. However, the task is presented as a knowledge-base completion task, and the token-based models may not be the best choice for this task.
Meta-Reviewer	Okay, I understand. What about the related work on set-based models? Is there any extension of the related work in the paper?
Dialogue Agent	The authors have extended the related work on set-based models in their paper. They have presented some new approaches and compared them with the baseline models. However, the related work on set-based models should be extended further.
Meta-Reviewer	I see. Thank you for the information. Based on your provided information, I would rate this paper as a 6 out of 10, with a confidence level of 4. The paper has some interesting ideas and the authors have presented some examples to show the effectiveness of their approach. However, there are some concerns regarding the experimental setup and the token-based models. The related work on set-based models should be extended further.
Dialogue Agent	Thank you for your feedback.

Table 14: Generated Meta-Reviewing Dialogue from our dataset based on the provided reviews.

A.22.2 Debates

The topic of the paper is: 'Global Warming is Not a Crisis'. The arguments for for and against sides of the debate are provided below:

Against side: "Global warming is not a crisis. Please keep in mind what the proposition is. It is not a debate over whether the earth has been warming over the past century. Uh, the earth is always warming or cooling, at least a few tenths of a degree. And we're talking about, so far, something on the order of six tenths of a degree centigrade. We're not even arguing about whether greenhouse gas emissions are contributing at some level to warming. And they most certainly should or I would suggest it would be very little. Indeed, as far as I can tell, even our opponents do not claim that global warming is a crisis at present. Rather, we are primarily addressing the future. Now, much of the current alarm, I would suggest, is based on ignorance of what is normal for weather and climate. [...]

For side: "The motion before us, global warming is not a crisis, means we ought to know what crisis means. The word does not mean catastrophe or alarmism. It means a crucial or decisive moment, a turning point, a state of affairs in which a decisive change for better or worse is imminent. We are talking about the future here. The entire world now really does have a critical choice to make. It is whether to continue on the present path of adding more and more carbon dioxide and other greenhouse gases to the atmosphere or whether to find another path. We're speaking of the future. And science tells us that the path we choose will largely determine what kind of earth our children and grandchildren will inherit. Our task tonight is to persuade you that global warming is indeed a crisis in exactly that precise sense so you should vote against the motion. The science community today has impeccable settled science, despite what you have just heard, that demonstrates the reality of global warming and its primary origin in human activities. [...]

Debate Decision-Making Dialogue	
Decision Maker	Hello, I am here to learn more about the debate topic, “Global warming is not a crisis." Can you provide me with information about both sides of the argument?
Dialogue Agent	Of course, for the “For" side, they argue that global warming is not a crisis at present and much of the current alarm is based on ignorance of what is normal for weather and climate. They suggest that extreme weather events always occur without evidence of systematic increases. Additionally, the sea level has been increasing since the end of the last Ice Age glaciation. The impact of warming on agriculture and disease seems dubious at best. On the other side, the “Against" arguments state that global warming is indeed a crisis in the sense that it is a turning point, a state of affairs in which a decisive change for better or worse is imminent. They argue that science has proven the reality of global warming and that mankind’s fingerprints have now clearly emerged above the noise of natural variability. They also mention the increasing frequency of heat waves, higher sea levels, and vanishing glaciers as evidence of the severity of the situation.
Decision Maker	Thank you for providing that information. Can you tell me about any projections or forecasts related to global warming?
Dialogue Agent	Yes, the "Against" side states that they project further warming of half a degree Fahrenheit for the next twenty-five years. Beyond that, it largely depends on how much more CO2 and other greenhouse gases humanity dumps into the atmosphere. They also warn that continuing to fuel the world using present technology will bring dangerous and possibly surprising climate changes by the end of this century, if not sooner.
Decision Maker	What is the scientific community’s stance on global warming?
Dialogue Agent	The "Against" side argues that today’s science community has impeccably settled science that demonstrates the reality of global warming and its primary origin in human activities. They also mention the Intergovernmental Panel on Climate Change’s report and the extensive process of experts who represent mainstream science. On the other hand, the “For" side acknowledges that even their opponents do not claim that global warming is a crisis at present and suggest that the impact of CO2 per unit of CO2 actually goes down with increasing CO2. They also argue that it is crucial to distinguish between the claim that models can display past behavior from the actual situation, which is that models can be adjusted to display past behavior once that behavior is known.
Decision Maker	Thank you for the information. Based on both sides of the argument, what do you suggest is the conclusion?
Dialogue Agent	As a dialogue agent, I cannot express opinions or make decisions. However, I provided the arguments from both sides that may help you reach your own conclusion.

Table 15: Debate Dialogue on the topic ‘Global warming is not a crisis’ generated from the provided arguments.

A.22.3 Product Reviews

The product in discussion is “Flexion KS-902 Kinetic Series Wireless Bluetooth Noise Cancelling Headphones with Microphone, Black". The reviews are below:

Reviews: “The battery life is sufficient where I only need to charge once every 2 days These headsets are equipped with the new Bluetooth 9 edition too advance for current music, so don’t even bother to enable because they don’t work. \n

These gave very good sound. Good sound quality, fits pretty well in your ear And it is also bluetooth, so I can connect it to my phone easily!\n

The user interface is really bad, when I try to turn it off, it usually calls someone on my phone, so I have had several butt dials thanks to these earphones.\n

There well rounded in all areas something that others don’t have. And they totally out perform the 2 hour battery life of the Flexion. The case is badly designed and too small.\n

These work really great when you want to listen to something without bothering other people. sleep.- Bass is amazing and really good- Loud and really clear For the price I paid they work excellent!\n

my background noise was amplified. He told me that he loves the headphones, but it was the amazing C.S that will have him as a returning customer. I recommend it to everyone! They actually turn on with a quick click, not a 3 second hold as in the instructions. The sound is just fantastic, the best sound I ever hear from an ear bud type speaker. Then my replacement pair broke, the cord just separated from the ear bud during normal use (see picture attached). \n

Sound quality is good and has buttons on the ear pieces for volume control, answering phone calls, and changing songs! The pause button and the advance track button doesn’t work and the previous track button is awkward and works only some of the time. As a commuter using public transportation, I was looking for something that would eliminate wires going to to my device and these do a great job of that. build quality is great compared to the price they do look expensive* I love that they were really easy to sync, which is a huge bonus for me! The headphones connected to my iPhone with no problems. \n

Sound quality was great but even after a full night of charging the batteries died after 2 hours. they give you a low battery warning a few seconds before they shut off which is kinda pointless. \n [...]

Product Buying Dialogue	
Buyer	Hi, I'm interested in purchasing the Flexion KS-902 Kinetic Series Wireless Bluetooth Noise Cancelling Headphones with Microphone, Black. Can you provide me with some information about the product?
Dialogue Agent	Sure, what would you like to know?
Buyer	Can you tell me about the battery life of the headphones?
Dialogue Agent	According to the reviews, the battery life is sufficient where you only need to charge once every 2 days. However, there have been a few reviewers who experienced a shorter battery life.
Buyer	What about the sound quality?
Dialogue Agent	The reviews state that the sound quality is good, and they worked really great when you want to listen to something without bothering other people. However, there were some reviews that reported the sound quality wasn't that great, and some experienced full of static if you get more than 7 feet from your phone.
Buyer	Is it easy to connect to my phone?
Dialogue Agent	Yes, the headphones are equipped with Bluetooth and are easy to sync with your iPhone or Android device.
Buyer	What about the user interface? Is it easy to use?
Dialogue Agent	According to some reviews, the user interface is really bad. They experienced several butt dials when trying to turn it off. However, others found the headphones to be easy to connect and use.
Buyer	With all that information, do you recommend these headphones?
Dialogue Agent	I'm sorry, I cannot provide recommendations or opinions. Please use the reviews provided to make an informed decision.
Buyer	Okay, thanks for the information. I think I'll go ahead and purchase them!

Table 16: Dialogue generated for the product 'Flexion KS-902 Kinetic Series Wireless Bluetooth Noise Cancelling Headphones with Microphone, Black' with the reviews provided before.