Adaptive Multi-Agent Response Refinement in Conversational Systems

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable success in conversational systems by generating human-like responses. However, they can fall short, especially when required to account for personas or specific knowledge. In real-life settings, it is impractical to rely on users to detect these errors and request a new response. One way to address this problem is to refine the response before returning it to the user. While existing approaches focus on refining responses within a single LLM, this method struggles to consider the diverse aspects needed for effective conversations. In this work, we propose refining LLM responses through a multi-agent framework, where each agent is assigned a specific role in the conversation. We focus on three key aspects crucial to conversational quality: factuality, personality, and coherence. Each agent is responsible for reviewing and refining one of these aspects, and their feedback is then merged to improve the overall response. To enhance collaboration among the agents, we introduce a dynamic communication strategy. Instead of following a fixed sequence of agents, our approach adaptively selects and coordinates the most relevant agents based on the specific requirements of each query. We validate our framework on several challenging conversational datasets, demonstrating that our multi-agent approach significantly outperforms relevant baselines, particularly in tasks involving knowledge or persona integration, or both.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable performance across a broad spectrum of NLP tasks, primarily due to their ability to generate coherent and contextually relevant responses, powered by extensive training on diverse data (Liu et al., 2024b; Dong et al., 2024). However, LLMs do not always produce satisfactory responses on the first attempt (Chiesurin et al., 2023; Lee et al., 2023a), and this issue becomes more evident in multi-turn conversational settings, where models must not only interpret the user's current query but also consider the entire conversational history, including dependencies, ambiguities, and co-references (Kim et al., 2021; Jang and Lukasiewicz, 2023; Gan et al., 2024). These difficulties are particularly pronounced in complex, personalized conversations that demand both persona alignment and factual accuracy (Semnani et al., 2023; Jandaghi et al., 2024). In such situations, users frequently have to prompt the LLM to correct its responses, which can interrupt the conversation flow and diminish the overall user experience (Madaan et al., 2023; Deng et al., 2023; Kim et al., 2024a). Thus, LLMs should proactively refine their inaccurate responses, even without explicit user requests for correction.

Several approaches have been explored for refining responses using a single agent, such as Self-Refine (Madaan et al., 2023), where a single agent handles an entire refinement process, generating feedback across multiple aspects and iteratively improving its outputs based on that feedback. However, relying solely on a single agent for refinement can be suboptimal, as the agent may become overly confident in the initial output, leading to bias (Liang et al., 2023), which can be amplified through repeated iterations of refinement (Xu et al., 2024b; Huang et al., 2024). This issue could be particularly critical for multi-turn conversational tasks, where errors in earlier turns can propagate to subsequent turns (Jeong et al., 2023).

A more effective approach is to incorporate multiple agents, each with a specific perspective. This method takes advantage of LLMs' ability to tackle complex tasks by decomposing them into simpler subtasks and enabling collaborative problem-solving (Wu et al., 2024; Guo et al., 2024; Du et al., 2024). Additionally, the multi-agent framework enables for integration of a broader range of tools and

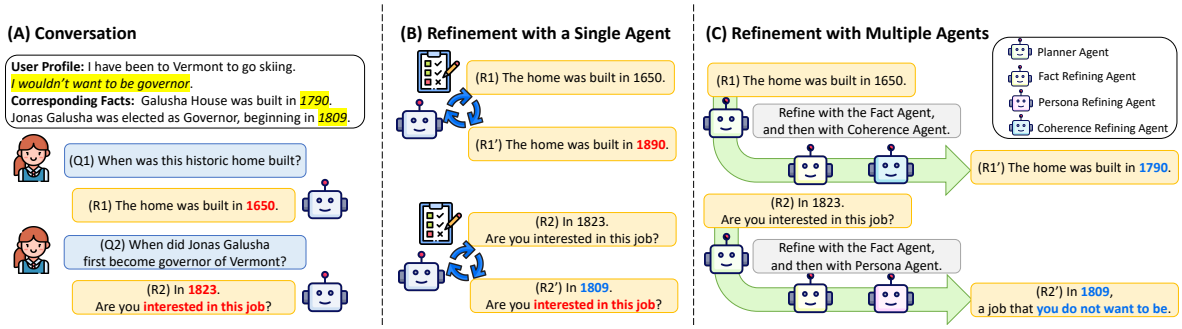

Figure 1: (A) A customized conversation which requires alignment with both the user profile and specific fact. The responding agent fails to provide accurate information in R1 and also fails to align with the user's profile in R2, as indicated in red. (B) In a single-agent-based refinement, one agent manages all the refinement procedures. While some misalignments can be addressed, as indicated in blue, the result may still be suboptimal. (C) Our multi-agent-based refinement, where multiple agents, each focusing on a different aspect, collaboratively refine responses based on the set and sequence determined by a planner agent.

knowledge sources (Shen et al., 2024; Gao et al., 2024; Xi et al., 2023), ensuring that each agent can be specifically tailored to meet the diverse needs of each task. While multi-agent frameworks have demonstrated great promise in fields such as human behavior simulation, economic theories, and more, their potential for response refinement in conversational systems remains largely underexplored. Figure 1 (A) exemplifies a conversation where the assistant responses should accurately reflect the user profile and specific knowledge. In such setting, employing multiple refining agents with specialized perspectives can significantly enhance conversation quality. Therefore, in this work, we introduce multiple refining agents that refine responses when the initial outputs are incorrect, each focused on a distinct perspective: persona alignment (persona-refining agent), factual grounding to mitigate hallucinations (fact-refining agent), and logical coherence with previous conversational turns (coherence-refining agent).

The remaining challenge, then, is how to enhance the overall quality of the conversation to deliver the most satisfactory refined response to the user, utilizing our three refining agents. To achieve this, the agents should collaborate differently for each query, as each query, even within the same conversation, may require a different focus. For example, as shown in Figure 1 (A), some queries require factual knowledge, while some queries also demand an understanding of the user's profile. Therefore, the set of agents deployed should vary depending on the query's specific needs. Additionally, the optimal sequence in which the agents are engaged may vary, as the focus and priorities of each query can differ. For example, a query requiring immediate factual verification might first engage the fact-refining agent to correct inaccuracies, followed by other agents as necessary. Therefore, we further propose a dynamic refinement process that adjusts both the set and the sequence of refining agents for each query, by introducing a novel planner agent that generates a sequence of required refining agents tailored to the query, along with justifications for each decision. Each refining agent then refines the response sequentially, following the sequence and referencing the justifications provided by the planner agent. Note that all agents are based on unsupervised LLMs, each instantiated with a prompt specifying a specific role. We refer to this framework as **M**ulti-**A**gent **R**efinement with **A**daptive agent selection (MARA), which is illustrated in Figure 1 (C).

We validate the efficacy of our framework on three challenging conversational datasets that require specific knowledge, INSCIT (Wu et al., 2023), or persona integration, PersonaChat (Zhang et al., 2018), and both of them, FoCus (Jang et al., 2022). The results show that MARA significantly improves performance over relevant baselines.

Our contributions and findings are threefold:

- We propose to use multiple agents for refining responses in conversational scenarios: fact-refining, persona-refining, and coherence-refining agents.
- We introduce a dynamic strategy with a planner agent that assigns and sequences refining agents based on the specific needs of each query.
- We demonstrate the efficacy of our framework across diverse conversational scenarios.

2 Related Work

LLMs in conversation systems Recent Large Language Models (LLMs) have demonstrated remarkable performance across a range of NLP tasks (OpenAI, 2023; Anthropic, 2024; Touvron et al., 2023), even without additional training, largely due to their massive pretraining on di-

verse datasets and their human-like reasoning abilities (Min et al., 2022; Wei et al., 2022). Their success is also remarkable in multi-turn conversational tasks (Chae et al., 2023; Liu et al., 2024b; Dong et al., 2024), as LLMs are able to generate human-like responses through effective prompting strategies (Lee et al., 2023b; Wadhwa et al., 2024; Deng et al., 2023). However, LLMs occasionally produce unsatisfactory responses, particularly when they fail to consider a user's profile (Jandaghi et al., 2024), specific knowledge (Semnani et al., 2023), or not coherent with the previous conversational turns (Huang et al., 2020; Ou et al., 2024).

Verification and Refinement with LLMs To address the issue of unsatisfactory responses, several studies have explored diverse approaches for verifying and refining incorrect outputs using LLMs. While some research involves the use of additional models to perform this verification and refinement, (Baek et al., 2023; Xu et al., 2024a), more recent studies emphasize the capability of LLMs to verify and refine their own responses without the need for further training (Shinn et al., 2023; Yao et al., 2023; Liu et al., 2024a; Gou et al., 2024; Chen et al., 2024a). Specifically, Madaan et al. (2023) proposed a method where a single agent LLM generates feedback on multiple aspects of its own response and iteratively refines the output by incorporating that feedback. In addition, Wang et al. (2024) introduced a self-collaboration approach, where multiple self-generated personas work collaboratively to solve a task using a single prompt. However, these single-agent-based refinements can be less effective, as the overall performance is constrained by the capacity of the agent, and once an agent gains confidence in its outputs, it may struggle to further generate novel thoughts, which can introduce bias (Liang et al., 2023), and this bias may be amplified during iterative refinement (Xu et al., 2024b; Huang et al., 2024).

Multi-agent LLMs Along with the powerful reasoning capabilities of an LLM, its performance can be further enhanced when multiple LLMs are involved by collaborating or debating each other, particularly on complex problems that require diverse perspectives from specialized LLM agents (Li et al., 2023; Wu et al., 2024; Guo et al., 2024; Rasal, 2024). Multi-agent LLM framework has recently been widely applied to various tasks, including software development (Hong et al., 2024; Qian et al., 2024), model evaluation (Chan et al., 2024), research topic generation (Baek et al., 2024), diagnostic consultations (Cao, 2023; Kim et al., 2024c), recommendation (Fang et al., 2024; Spurlock et al., 2024), and reasoning improvement (Chen et al., 2024b), which is a concurrent work to ours. Specifically, Kim et al. (2024b) proposed refining an LLM's responses to generate more accurate fact-checking explanations by employing two debating agents for a single-turn task. In contrast, we address the more challenging multi-turn conversational setting, which requires both persona alignment and factual accuracy, with each agent playing a specialized role in refining the conversational response from three crucial perspectives for customized conversations.

3 Method

We introduce MARA, a multi-agent refinement framework designed to dynamically refine responses in a multi-turn conversational setting.

3.1 Preliminaries

We begin by formally defining a Large Language Model (LLM) and the roles of responding in a multi-turn conversational setting.

Large Language Model Let us first define LLM as $\texttt{LLM}$, which takes an input sequence of tokens $\boldsymbol{x}$ and generates an output sequence of tokens $\boldsymbol{y}$. This process can be represented as $\boldsymbol{y} = \texttt{LLM}(\mathcal{P}(\boldsymbol{x}))$, where the prompt template $\mathcal{P}$ incorporates additional context or instructions that guide the LLM's behavior. Specifically, $\mathcal{P}$ can be used for role assignment, ensuring that the LLM adopts a particular role or performs a specific task.

Responding Agent in Conversation In a multi-turn conversational setting, the LLM, acting as a responding agent, generates a response $\boldsymbol{r}^i$ for the i-th turn based on the current query $\boldsymbol{q}^i$ and the preceding conversational context, which can be represented as: $\boldsymbol{r}^i = \texttt{LLM}_{\texttt{respond}}(\boldsymbol{q}^i, \boldsymbol{q}^{i-1}, \boldsymbol{r}^{i-1}, ..., \boldsymbol{q}^1, \boldsymbol{r}^1)$, where $\boldsymbol{q}^i$ is the user's query at the i-th turn, and $\boldsymbol{r}^i$ is the response generated by the $\texttt{LLM}$. However, initial responses from the LLM may be suboptimal, potentially overlooking the user's profile, specific knowledge, or failing to maintain coherence, and we should note that the errors in earlier turns can negatively impact subsequent turns.

3.2 Multi-Agent Response Refinement

We now turn to our primary focus of further refining the initial response. To achieve this, we define three specialized refining agents, each responsible for enhancing different aspects of the response.

Refining Agents We define a refining agent as $\texttt{LLM}_{\texttt{refine}}$, where $\texttt{LLM}$ is tasked with refining a given response. Specifically, we introduce three refining agents, each addressing a different aspect of the response:a fact-refining agent ($\texttt{LLM}_{\texttt{fact}}$) focusing on factual correctness, a persona-refining agent ($\texttt{LLM}_{\texttt{persona}}$) focusing on persona alignment, and a coherence-refining agent ($\texttt{LLM}_{\texttt{coherence}}$) focusing on response coherence. Each agent is instantiated with its respective prompt template: $\mathcal{P}_{\texttt{fact}}$, $\mathcal{P}_{\texttt{persona}}$, and $\mathcal{P}_{\texttt{coherence}}$. Note that, unlike a single-agent refinement systems, a multi-agent framework allows each refining agent to access distinct operations, tools, or LLMs. For instance, the fact-refining agent may use a larger or more up-to-date LLM to correct hallucinations, while the persona-refining agent can access a persona pool to improve alignment.

Refining Process for Refining Agents Then, how does the refining process proceed for each refining agent? For any given response $\boldsymbol{r}$, a refining agent operates through a step-by-step process, beginning by verifying the alignment of the response with its assigned feature and generating justifications for this verification. The agent then refines the response as necessary and produces justifications for any changes made. Therefore, the output of the refinement process is represented as: $\boldsymbol{o}_{\texttt{refine}} = [\boldsymbol{r}_{\texttt{verify}}, \boldsymbol{j}_{\texttt{verify}}, \boldsymbol{r}_{\texttt{refine}}, \boldsymbol{j}_{\texttt{refine}}]$, where $[\cdot]$ is a concatenation operation, $\boldsymbol{r}_{\texttt{verify}}$ is a decision indicating whether the original response aligns with the refining agent's target feature, $\boldsymbol{j}_{\texttt{verify}}$ is the justification for this verification decision, $\boldsymbol{r}_{\texttt{refine}}$ is the refined response, and $\boldsymbol{j}_{\texttt{refine}}$ is the justification for the refinements made to the given response. Furthermore, to ensure that each refining agent is aware of its prior refinements within the same conversation, we incorporate the outputs of the agent's previous refinements as context for its current operation. Therefore, the refining process at turn i is represented as: $\boldsymbol{o}^{i}_{\texttt{refine}} = \texttt{LLM}_{\texttt{refine}}(\boldsymbol{r}^{i}, [\boldsymbol{o}^{j}_{\texttt{refine}}]_{j \in \mathcal{K}^{i}})$, where $\mathcal{K}^{i}$ is the set of previous turns where refinement occurred.

3.3 Communication Strategy among Agents

To enable effective collaboration among the agents, we introduce various communication strategies involving our refining agents and propose a dynamic strategy to optimize the refinement process.

Simultaneous Communication As an initial approach to collaborative communication among the three refining agents, we introduce a simultaneous communication strategy. In this approach, each refining agent independently verifies and refines the initial response based on its designated role. The outputs from these agents are then passed to a finalizer agent ($\texttt{LLM}_{\texttt{finalizer}}$), which aggregates the individual refinements and feedback into a single unified response. Formally, this process can be expressed as: $\texttt{LLM}_{\texttt{finalizer}}(\boldsymbol{o}_{\texttt{fact}}, \boldsymbol{o}_{\texttt{persona}}, \boldsymbol{o}_{\texttt{coherence}})$. However, this simultaneous approach always necessitates an additional finalizer agent, and the overall quality of the final output may heavily depend on the capabilities of this finalizer agent.

Sequential Communication To address potential limitations of the simultaneous approach, we introduce a sequential refinement process where each agent builds upon the response refined by the previous agent. Specifically, given an initial response $\boldsymbol{r}$, the response is sequentially refined by a series of agents denoted as $\boldsymbol{s} = [\texttt{LLM}_{\texttt{refine}_k}]_{k=1}^{n}$, where each agent $\texttt{LLM}_{\texttt{refine}_k}$ is chosen from the set of three refining agents, $\{\texttt{LLM}_{\texttt{fact}}, \texttt{LLM}_{\texttt{persona}}, \texttt{LLM}_{\texttt{coherence}}\}$. Here, each refining agent $\texttt{LLM}_{\texttt{refine}_k}$ receives the response refined by the preceding agent, denoted as $\boldsymbol{r}_{\texttt{refine}_{k-1}}$ as input, and uses its specific prompt template $\mathcal{P}_{\texttt{refine}_k}$ to produce the next refined response. The sequential process continues through the sequence of agents until the final refined response $\boldsymbol{r}_{\texttt{refine}_n}$ is produced. However, it is may not be feasible to predefine a fixed sequence of refining agents, as each query $\boldsymbol{q}$ may require a different combination and order of refining agents based on the specific context. For instance, one query may require only fact-checking, while another may benefit from persona alignment. In some cases, a query might need both, but the sequence in which the refining agents are engaged can vary depending on the conversational context.

Dynamic Sequential Communication To further optimize the response refinement process, we introduce a dynamic strategy where a planner agent selects what it considers to be the most

suitable sequence of refining agents, adapting to the needs of each query. Specifically, the planner agent, denoted as $\texttt{LLM}_{\texttt{planner}}$, organizes both the set and sequence of refining agents required for each query, instantiated by a template $\mathcal{P}_{\texttt{planner}}$. Given a query $\boldsymbol{q}$ and an initial response $\boldsymbol{r}$, the planner agent outputs the sequence of refining agents $\boldsymbol{s}_{\texttt{planner}}$, along with justifications for the selected agents and their order, $\boldsymbol{j}_{\texttt{planner}}$, as follows: $\boldsymbol{o}_{\texttt{planner}} = \texttt{LLM}_{\texttt{planner}}(\boldsymbol{q}, \boldsymbol{r})$, where $\boldsymbol{o}_{\texttt{planner}} = [\boldsymbol{s}_{\texttt{planner}}, \boldsymbol{j}_{\texttt{planner}}]$. Once $\boldsymbol{s}_{\texttt{planner}}$ is determined, the refining agents sequentially refine the response, with each agent taking as input the refined response from the preceding agent in the sequence. Additionally, each agent also receives $\boldsymbol{j}_{\texttt{planner}}$, which contains the planner's justifications, allowing each agent to understand its role in the sequence and collaborate effectively.

4 Experimental Setups

In this section, we describe the experimental setup, leaving further details in Appendix A.

4.1 Datasets

We validate MARA in a realistic conversational setting by testing it on three conversational datasets, including persona-aligned and knowledge-grounded queries. **1) FoCus** (Jang et al., 2022) consists of conversational queries requiring both persona-aligned and knowledge-grounded responses, with knowledge sourced from Wikipedia. **2) PersonaChat** (Zhang et al., 2018) is a dataset focusing on the persona aligned conversation. **3) INSCIT** (Wu et al., 2023) is an information-seeking conversational dataset, where the knowledge is grounded in Wikipedia.

4.2 Models

We compare MARA against four baseline models, starting with a baseline without refinement, two single-agent refinement approaches, and one multi-agent refinement approach. **1) No Refine** is an approach involves generating a response solely with a responding agent, without any further refinement of the response. **2) Self-Refine** (Madaan et al., 2023) is a representative single agent refinement baseline, which makes an LLM to generate feedback on 10 aspects (e.g., Relevant, Informative, Engaging) of its own response and iteratively refines the output by incorporating this feedback. **3) SPP** (Wang et al., 2024) is another single-agent refinement baseline where a single agent generates multiple self-generated personas and makes them collaborate to solve a task using a single prompt. **4) MADR** (Kim et al., 2024b) is a multi agent refinement baseline where two agents debate based on the predefined error types to generate fact checking explanations. They iteratively debate until the judge agent decides that two agents are in consensus, and the finalizer agent refines the refined response based on the feedback generated during a debate. **5) MARA** is our proposed system, where the agents refine the response in the specified order, adaptively generated by a planner agent.

4.3 Evaluation Metrics

We evaluate the models using G-Eval (Liu et al., 2023) to assess the quality of the refined responses. Following the setup in Liu et al. (2023) for a conversational setting, we assess each refined response using four metrics: **1) Coherence (Coh.)** measures whether the conversation response logically follows the preceding context with a scale of 1 to 3. **2) Groundedness (Grd.)** evaluates whether the response accurately incorporates the provided fact, with a scale of 0 to 1. **3) Naturalness (Nat.)** evaluates whether the response is natural with a scale of 1 to 3. **4) Engagingness (Eng.)** measures whether the response is engaging with a scale of 1 to 3.

4.4 Implementation Details

We use Claude Sonnet 3 model (Anthropic, 2024) as a base LLM as a responding agent. For the refining agents, we employ the Claude Sonnet 3 model for the persona-refining, coherence-refining, and planner agents. Since a multi-agent framework offers the flexibility to assign different LLMs or tools to each agent based on their role, for the fact-refining agent, we use the more advanced Claude Sonnet 3.5 model, as it offers improved capabilities in factual accuracy and reasoning, which are critical for this particular role. Regarding the datasets, we sample 100 conversations for each dataset. For the G-Eval metric, we use GPT-4o mini model (OpenAI, 2023), and follow the implementation setup in the G-Eval paper (Liu et al., 2023), including the normalization steps for each evaluation per query to mitigate potential bias. We include the prompts used for MARA and G-Eval in the Appendix.

5 Experimental Results and Analyses

We now present the overall experimental results and provide detailed analyses.

Table 1: Results on three datasets, using Claude as the base LLM. The significantly best results are highlighted in **bold**.

Methods	FoCus				PersonaChat				INSCIT			
	Coh. (1-3)	Grd. (0-1)	Nat. (1-3)	Eng. (1-3)	Coh. (1-3)	Grd. (0-1)	Nat. (1-3)	Eng. (1-3)	Coh. (1-3)	Grd. (0-1)	Nat. (1-3)	Eng. (1-3)
No Refine	2.39 ± 0.02	0.49 ± 0.02	2.02 ± 0.01	2.15 ± 0.02	1.87 ± 0.01	0.42 ± 0.00	1.55 ± 0.01	1.55 ± 0.00	2.53 ± 0.02	0.52 ± 0.02	**2.36** ± 0.01	2.36 ± 0.01
Self-Refine	2.10 ± 0.01	0.37 ± 0.01	1.87 ± 0.00	2.05 ± 0.02	2.26 ± 0.01	0.52 ± 0.01	**2.07** ± 0.01	2.30 ± 0.03	2.31 ± 0.01	0.45 ± 0.00	2.19 ± 0.01	2.24 ± 0.01
SPP	2.44 ± 0.00	0.47 ± 0.01	1.99 ± 0.02	2.46 ± 0.02	1.98 ± 0.02	0.44 ± 0.01	1.72 ± 0.02	2.07 ± 0.01	2.52 ± 0.01	0.49 ± 0.01	2.29 ± 0.01	2.47 ± 0.02
MADR	1.93 ± 0.03	0.32 ± 0.01	1.68 ± 0.04	1.54 ± 0.03	1.59 ± 0.03	0.28 ± 0.01	1.38 ± 0.03	1.32 ± 0.01	2.29 ± 0.03	0.46 ± 0.01	2.12 ± 0.04	1.89 ± 0.04
MARA (Ours)	**2.67** ± 0.00	**0.65** ± 0.01	**2.15** ± 0.01	**2.83** ± 0.00	**2.32** ± 0.01	**0.56** ± 0.01	1.98 ± 0.01	**2.54** ± 0.06	**2.62** ± 0.01	**0.59** ± 0.01	2.26 ± 0.04	**2.73** ± 0.06

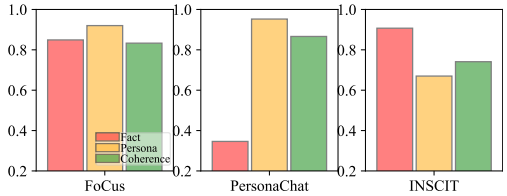

Figure 2: Distribution of refining agents for three datasets.

Planner Design	Coh.	Grd.	Nat.	Eng.	#Agent
No Refine	2.21	0.40	1.88	2.05	1.0
MARA w/ Random	2.42	0.47	1.95	2.64	2.0
MARA	2.54	0.58	2.07	2.76	4.4
MARA w/ Ideal	**2.88**	**0.80**	**2.58**	**2.93**	3.0

Table 2: Analyses on a planner agent, comparing performance with randomly or ideally assigned refining agents.

5.1 Main Results

Here, we present the overall results across various conversational scenarios.

Overall Results We report the overall experimental results in Table 1 with three different runs. MARA consistently outperforms other models across divserse metrics, and the gaps between MARA and the other models are both substantial and statistically significant. To evaluate the statistical significance of these differences, we performed Tukey's Honest Significant Difference (HSD) test, detailed in Appendix B.

Specifically, in terms of coherence on FoCus, MARA shows a notable improvement over the MADR model, with a mean difference of 0.7456, indicating a substantial gap in favor of MARA. Similarly, MARA performs significantly better than the Self-Refine model, with a gap of 0.5691, and also outperforms SPP by 0.2317. In the groundedness metric, MARA continues to demonstrate significant improvements, outperforming MADR by 0.3337 and Self-Refine by 0.2806, while showing a meaningful advantage over SPP with a gap of 0.1796. These results highlight MARA's ability to generate more factually grounded and coherent responses. When it comes to naturalness, MARA exhibits a strong advantage over MADR with a mean difference of 0.4772, and a similarly significant improvement over Self-Refine, with a gap of 0.2851. Compared to SPP, MARA performs better by 0.1616. In terms of engagingness, MARA shows the largest gap in performance, particularly over MADR, with a mean difference of 1.2939, demonstrating its capacity to produce significantly more engaging interactions. It also outperforms Self-Refine by 0.7801 and SPP by 0.3706.

Analyses on Diverse Conversational Settings Furthermore, we explore diverse conversational scenarios, particularly challenging ones requiring persona alignment (PersonaChat), factual grounding (INSCIT), and a combination of both (FoCus). In the PersonaChat dataset, which emphasizes responses aligns the user's profile, single-agent refinement methods, including Self-Refine and SPP, improve over the baseline without refinement. This is likely because these methods focus on improving fluent flow of generation, which aligns well with persona-based tasks that require maintaining consistency in the user's tone and preferences across the conversation rather than managing specific factual content. However, on datasets requiring specific knowledge, such as INSCIT and FoCus, single-agent refinement becomes less effective and even leads to performance degradation, particularly in groundedness scores. Interestingly, SPP outperforms Self-Refine, suggesting that generating multiple personas with a single agent can be beneficial for information-intensive queries. Nevertheless, compared to the significantly improved performance of our multi-agent framework, generating multiple perspectives with distinct agents appears to be more effective, and single-agent persona generation remains suboptimal.

While the multi-agent baseline, MADR, refines factual errors through iterative debates between two agents based on predefined error typologies, it still struggles to effectively improve upon the baseline without refinement, even for groundedness. This suggests that the predefined error types may not be sufficiently generalizable to address the broader range of issues encountered in diverse conversational contexts. Furthermore, the use of the same LLM for refining factual errors as the responding

Table 3: Results from human evaluation and Spearman's correlation with G-Eval on the same subset of the FoCus dataset.

Methods	Coh.	Grd.	Nat.	Eng.
No Refine	2.53	0.57	2.47	2.10
Self-Refine	2.04	0.36	1.86	1.93
SPP	2.49	0.53	2.44	2.18
MADR	2.03	0.35	2.10	1.65
MARA (Ours)	**2.69**	**0.79**	**2.61**	**2.75**
Spearman's ρ	0.51	0.48	0.35	0.58

Table 4: Ablation studies on each refining agent in FoCus, including incorporation of all perspectives into a single agent.

Refining Agent	Coh.	Grd.	Nat.	Eng.
No Refine	2.39	0.49	2.02	2.15
w/ Fact	2.62	0.61	2.14	2.52
w/ Persona	2.54	0.49	**2.18**	2.77
w/ Coherence	2.46	0.53	1.89	2.57
w/ Single	2.54	0.58	2.15	2.55
MARA (Ours)	**2.67**	**0.65**	2.15	**2.83**

agent might contribute to this problem, as it can fail to correctly verify factual inaccuracies and may reinforce the same hallucinated errors produced by the initial responding agent. In contrast, our MARA consistently enhances performance across all three conversational scenarios, indicating the effectiveness of our three-agent design with a more flexible choice of LLMs, focusing on persona alignment, factual accuracy, and coherence.

Analyses on Planner Agent Since our planner agent plays a crucial role in dynamically allocating the appropriate refining agents based on the given query and its initial response, we further investigate its effectiveness. First, we analyze the distribution of refining agents used across various conversational settings, as shown in Figure 2. Interestingly, the distribution varies across different conversational datasets. To be specific, the persona-refining agent is predominantly used in the persona-focused dataset, while the fact-refining agent is more frequently required in the knowledge-focused dataset. For the dataset requiring both aspects, both agents are allocated in similar proportions. This suggests that our planner agent operates effectively in adapting to different conversational requirements. Next, we further evaluate the effectiveness of the current design of our planner agent on the subset of the FoCus dataset with 20 conversations, as shown in Table 2. The performance of a randomly assigned agents is lower than ours with a planner, indicating that our planner effectively defines a sequence for the refining agents. However, even with the randomly assigned agents, it outperforms the baseline without refinement, showing the effectiveness of using our three refining agents. Finally, we explore the full potential of our planner agent by investigating how much further it can improve performance. To do so, we evaluate an ideal planner that selects the best-performing sequence. We report its performance by brute-forcing all possible combinations of sequences using our three refining agents and selecting the optimal sequence for each query. As shown in Table 2, this ideal planner achieves the highest performance while requiring fewer accesses to the refining agents. These results validate the effectiveness of our dynamic refining agent allocation strategy and indicate that further advancements in planner agents could yield even greater performance improvements.

5.2 Ablations and Discussion

In this section, we provide a detailed analysis of the performance improvements.

Human Evaluation Although G-Eval is a widely used model-based evaluation metric for its high correlation with humans, we further conduct a human evaluation to measure the alignment between their evaluations. Specifically, 8 English-fluent participants were involved in evaluating 38 conversations, covering a total of 288 queries. The results of the human evaluation, as shown in Table 3, highlight the effectiveness of our MARA.

Additionally, to validate the reliability of G-Eval, we further assess the alignment between G-Eval and human judgments by calculating Spearman's rank correlation between the model-based evaluation and human evaluation results. The correlation values for coherence (0.51) and groundedness (0.48) indicate a reasonable alignment between G-Eval and human rankings, suggesting that G-Eval is fairly reliable in assessing these metrics. For engagingness, the correlation of 0.58 demonstrates the strongest alignment, indicating that G-Eval is particularly effective at evaluating how engaging a conversation is. However, the relatively low correlation for naturalness (0.35) reveals a noticeable gap between G-Eval's assessments and human evaluations, especially in capturing the human-like qualities of the responses. This result might help explain the relatively smaller gaps in naturalness among the models, as shown in Table 1. Furthermore, we provide a qualitative case study in Table 11.

Table 5: Results with simultaneous (Sim.) and sequential (Seq.) communication strategies on the FoCus dataset.

Strategy	Coh.	Grd.	Nat.	Eng.	#Agent
Sim. (F+C+P)	2.57	0.59	1.93	2.75	5.0
Seq. (F→C→P)	2.54	0.56	1.99	2.85	4.0
Seq. (C→P→F)	2.65	0.60	2.15	2.66	4.0
Seq. (P→F→C)	2.56	0.59	1.96	2.68	4.0
MARA (Ours)	**2.67**	**0.65**	**2.15**	**2.83**	4.4

Table 6: Analyses of the fact-refining agent size and the responding agent size in the FoCus dataset.

Model Size	Coh.	Grd.	Nat.	Eng.
No Refine w/ Sonnet 3	2.39	0.49	2.02	2.15
MARA w/ Fact Sonnet 3	2.56	0.57	2.00	2.80
MARA w/ Fact Sonnet 3.5	**2.67**	**0.65**	**2.15**	**2.83**
No Refine w/ Sonnet 3.5	2.50	0.58	2.10	2.32
MARA w/ Sonnet 3.5	**2.87**	**0.78**	**2.55**	**2.91**

Table 7: GPT4o-mini and Llama 3.1 8B, 70B on FoCus.

	Methods	Coh.	Grd.	Nat.	Eng.
GPT	**No Refine**	2.45	0.47	2.27	1.98
	MARA (Ours)	**2.61**	**0.57**	**2.32**	**2.70**
Llama	**No Refine - 8B**	2.30	0.39	2.09	1.88
	MARA - 8B (Ours)	**2.46**	**0.48**	**2.17**	**2.24**
	No Refine - 70B	2.41	0.44	2.25	1.88
	MARA - 70B (Ours)	**2.60**	**0.55**	**2.31**	**2.67**

Ablation Studies on Refining Agents In order to see how each refining agent contributes to the overall performance, we perform ablation studies when adding three refining agents. As shown in Table 4, each refining agent plays crucial role in overall performance. Additionally, we assess the performance of combining all three aspects—persona alignment, factual accuracy, and coherence—within a single agent. The improved results, compared to the No Refine baseline, validate the effectiveness of our proposed three perspectives. However, the superior performance of MARA highlights the effectiveness of multi-agent refinement over single-agent refinement, highlighting the advantages of distributed specialization for each agent.

Ablation Studies on Communication Strategy In order to investigate the efficacy of the dynamic sequential communication strategy, we compare it against other possible strategies using our refining agents. As shown in Table 5, the simultaneous communication strategy requires more accesses to the agents per query, as it involves all three refining agents as well as the additional finalizer agent. We also evaluate the performance of three different sequential settings, where the agents refine the initial response in different orders. The results demonstrate that performance varies significantly depending on the order of the refining agents, which might not be feasible in realistic conversational settings, since the user has to predefine the optimal sequence for every query. On the other hand, our strategy adapts to each query by selecting the relevant sequence of refining agents, leading to more consistent and improved performance.

Analyses on Fact-Refining Agent A notable advantage of the multi-agent framework is its flexibility in utilizing different agent. In our case, we assign a larger model to the fact-refining agent, since a larger model typically has more extensive knowledge. As shown in Table 6, while MARA with a fact-refining agent of the same model size as the responding agent remarkably improves performance, the results show that using a larger fact-refining agent yields much better outcomes, likely because it reduces the replication of the same hallucination errors. This highlights a key strength of the multi-agent framework: the ability to utilize more advanced LLMs to specific roles, thereby enhancing performance beyond what a single-agent approach can achieve. Then, some might ask why not simply use the larger model as the responding agent directly. However, the results in Table 6 show that refinement with MARA largely improves performance, even when the responding agent is already a larger model.

Effectiveness for Diverse LLMs To assess whether our proposed framework works effectively with other widely used LLMs, we validate it using two other LLMs, as shown in Table 7. The results show that our method can effectively refine initial responses across different LLMs, highlighting the applicability of MARA to diverse LLMs.

6 Conclusion

While LLMs have shown promise in conversational systems, they often struggle with complex, customized conversations requiring persona alignment and specific knowledge. Prior approaches using single-agent self-refinement can be suboptimal, as one model must handle all aspects of feedback and refinement. In this work, we presented a cooperative multi-agent framework to address these limitations, with specialized agents for fact-checking, persona alignment, and coherence, by allocating them dynamically tailored to each query. Our evaluation on challenging conversational datasets shows that this multi-agent framework significantly outperforms existing baselines.

Limitations

While our MARA framework demonstrates clear advantages by adaptively selecting the most suitable refining agents for diverse customized conversational settings, there remains room for improvement, particularly in the planner agent, as shown by the performance gap between the ideal planner and the current planner Table 2. Specifically, since our current planner is fully based on an unsupervised LLM, constructing a dataset with labeled ideal sets and sequences of required perspectives, followed by fine-tuning the planner on this data, could be a promising direction for future research.

While a multi-agent framework shows remarkable effectiveness, it may raise concerns about scalability and resource efficiency. Therefore, we further suggest some possible approaches that can address these challenges. As demonstrated by the significantly reduced number of LLM accesses with an optimized planner in Table 2, improving the planner agent can alleviate concerns regarding resource efficiency. To further enhance scalability, exploring lightweight or more efficient agent models would be also beneficial. In Table 6 and Table 7, we show that our proposed MARA performs well even with a smaller LLM. Notably, Table 7 shows significant improvements with a lightweight, open-sourced LLM, the Llama 3.1 8B model, which highlights its potential for scalability. Nonetheless, further exploration of scalability remains a promising avenue for future work, which we leave as a meaningful future work.

Additionally, further improvements on effectiveness of MARA could be achieved through iterative refinement of the initial response. However, it is important to note that this approach may require more frequent access to the refining agents, potentially reducing efficiency. Moreover, the flexibility of the multi-agent framework in selecting diverse external tools or LLMs opens up opportunities for incorporating tools such as Retrieval-Augmented Generation (RAG) systems to further enhance performance, which could also be interesting future research direction.

Ethics Statement

The extensive experimental results on MARA validate its applicability in realistic conversational scenarios, where a diverse range of queries with varying requests exists. However, given the potential diversity of real-world user inputs, and as a system designed to assist humans safely, it is crucial to consider scenarios where inputs or generated responses may be offensive or harmful (Shin et al., 2024; Dong et al., 2024). To address this, developing an additional agent to monitor both user inputs and generated content for harmful or offensive material is essential. We consider this a critical area for future research.

References

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

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arxiv Preprint arXiv:2404.07738*, abs/2404.07738.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-augmented language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1720–1736. Association for Computational Linguistics.

Lang Cao. 2023. Diaggpt: An llm-based chatbot with automatic topic management for task-oriented dialogue. *arxiv Preprint arXiv:2308.08043*, abs/2308.08043.

Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5606–5632. Association for Computational Linguistics.

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

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024a. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024b. Magicore: Multi-agent, iterative, coarse-to-fine refinement for reasoning.

Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 947–959. Association for Computational Linguistics.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10602–10621. Association for Computational Linguistics.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for LLM conversation safety: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6734–6747, Mexico City, Mexico. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arxiv Preprint arXiv:2402.01135*, abs/2402.01135.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1645–1665. ELRA and ICCL.

Dawei Gao, Zitao Li, Weirui Kuang, Xuchen Pan, Daoyuan Chen, Zhijian Ma, Bingchen Qian, Liuyi Yao, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. Agentscope: A flexible yet robust multi-agent platform. *arXiv Preprint arXiv.2402.14034*, abs/2402.14034.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arxiv Preprint arXiv:2402.01680*, abs/2402.01680.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for A multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3):21:1–21:32.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15245–15270. Association for Computational Linguistics.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15970–15985. Association for Computational Linguistics.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10803–10812. AAAI Press.

Soyeong Jeong, Jinheon Baek, Sung Ju Hwang, and Jong Park. 2023. Realistic conversational question answering with answer selection based on calibrated confidence and uncertainty measurement. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 477–490. Association for Computational Linguistics.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6130–6141. Association for Computational Linguistics.

Hana Kim, Kai Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. 2024a. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 104–123, St. Julian's, Malta. Association for Computational Linguistics.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024b. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arxiv Preprint arXiv:2402.07401*, abs/2402.07401.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024c. Mdagents: An adaptive collaboration of llms for medical decision-making. *Preprint*, arXiv:2404.15155.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023a. Platypus: Quick, cheap, and powerful refinement of llms. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023b. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4536–4554. Association for Computational Linguistics.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, abs/2305.19118.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, and Jinjun Xiong. 2024a. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*, abs/2406.15673.

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. 2024b. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arxiv Preprint arXiv:2401.02777*, abs/2401.02777.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, abs/2303.08774.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6137–6170. Association for Computational Linguistics.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15174–15186. Association for Computational Linguistics.

Sumedh Rasal. 2024. LLM harmony: Multi-agent communication for problem solving. *arXiv Preprint arXiv.2401.01312*, abs/2401.01312.

Sina J. Semnani, Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. 2023. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2387–2413. Association for Computational Linguistics.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024. Small llms are weak tool learners: A multi-llm agent. *arXiv Preprint arXiv.2401.07324*, abs/2401.07324.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask llms directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 16122–16143. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kyle Dylan Spurlock, Cagla Acun, Esin Saka, and Olfa Nasraoui. 2024. Chatgpt for conversational recommendation: Refining recommendations by reprompting with feedback. *ArXiv*, abs/2401.03605.

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

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to refine with fine-grained natural language feedback. *arXiv preprint arXiv:2407.02397*, abs/2407.02397.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 257–279. Association for Computational Linguistics.

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

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. In *COLM*.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Inscit: Information-seeking conversations with mixed-initiative interactions. *Trans. Assoc. Comput. Linguistics*, 11:453–468.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *arxiv Preprint arXiv:2309.07864*, abs/2309.07864.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024a. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1429–1445. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024b. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15474–15492. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics.

A Additional Experimental Setups

A.1 Datasets

We validate MARA in a realistic conversational setting by testing it on three conversational datasets, including persona-aligned and knowledge-grounded queries.

1) FoCus (Jang et al., 2022) consists of conversational queries requiring both persona-aligned and knowledge-grounded responses, with knowledge sourced from Wikipedia. The dataset has an average of 11.9 conversational turns. We use the publicly available validation set.

2) PersonaChat (Zhang et al., 2018) is a conversational dataset involving dialogues between two speakers, each having persona sentences that describe their character and serve as context for the conversation. The dataset has an average of 14.0 conversational turns. We use the publicly open validation set.

3) INSCIT (Wu et al., 2023) is an information-seeking conversational dataset, where the knowledge is grounded in Wikipedia. The dataset has an average of 11.8 conversational turns. We use the publicly accessible test set for the evaluation.

A.2 Models

We compare MARA against four baseline models, starting with a baseline without refinement, two single-agent refinement approaches, and one multi-agent refinement approach.

1) No Refine is an approach involves generating a response solely with a responding agent, without any further refinement of the response.

2) Self-Refine (Madaan et al., 2023) is a representative single agent refinement baseline, which makes an LLM to generate feedback on 10 aspects (e.g., Relevant, Informative, Engaging) of its own response and iteratively refines the output by incorporating this feedback.

3) SPP (Wang et al., 2024) is another single-agent refinement baseline where a single agent generates multiple self-generated personas and makes them collaborate to solve a task using a single prompt.

4) MADR (Kim et al., 2024b) is a multi agent refinement baseline where two agents debate based on the predefined error types to generate faithful fact checking explanations. They iteratively debate until the judge agent decides that two agents are in consensus, and the finalizer agent refines the refined response based on the feedback generated during a debate.

5) MARA is our proposed system, where a planner agent dynamically determines the optimal sequence of refining agents to involve, and then the agents refine the response in the specified order.

A.3 Evaluation Metrics

We evaluate the models using G-Eval (Liu et al., 2023) to assess the quality of the refined responses. Following the setup in Liu et al. (2023) for a conversational setting, we assess each refined response using four metrics:

1) Coherence (Coh.) measures whether the conversation response logically follows the preceding context with a scale of 1 to 3.

2) Groundedness (Grd.) evaluates whether the response is based on and accurately incorporates the provided fact, with a scale of 0 to 1.

3) Naturalness (Nat.) evaluates whether the response is natural, as if it were part of a real, human conversation, with a scale of 1 to 3.

4) Engagingness (Eng.) measures whether the response is engaging and interesting, with a scale of 1 to 3.

A.4 Implementation Details

We use Claude Sonnet 3 model (Anthropic, 2024) as a base LLM as a responding agent. Since a multi-agent framework offers the flexibility to assign different LLMs or tools to each agent based on their role, we employ the Claude Sonnet 3 model for the persona-refining, coherence-refining, and planner agents. For the fact-refining agent, we use the more advanced Claude Sonnet 3.5 model, as it offers improved capabilities in factual accuracy and reasoning, which are critical for this particular role. Regarding the datasets, we sample 100 conversations for each dataset. For the G-Eval metric, we use GPT-4o mini model (OpenAI, 2023), and follow the implementation setup introduced by Liu et al. (2023), including the normalization steps for each evaluation per query to mitigate potential bias. In order to further enhance the reliability of G-Eval, we also provide the references, such as the annotated gold response, corresponding fact, and persona, to the model. The prompts used for MARA are provided in Table 13 (responding agent), Table 14 (planner agent), Table 15 (fact-refining agent), Table 16 (persona-refining agent), and Table 17 (coherence-refining agent). The prompt templates for G-Eval are shown in Table 18, Table 19, Table 20, and Table 21.

Table 8: Significance Testing for Coherence, Groundedness, Naturalness, and Engagingness (FoCus Dataset)

	Comparison	Model 1	Model 2	Mean Diff.	P-adj	Lower	Upper	Significant?
Coherence	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.2838	0.000	-0.3332	-0.2344	Yes
	No Refine vs MADR	No Refine	MADR	-0.4619	0.000	-0.5153	-0.4084	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.2854	0.000	-0.3348	-0.2360	Yes
	No Refine vs SPP	No Refine	SPP	0.0521	0.0378	0.0027	0.1015	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	0.7456	0.000	0.6962	0.7950	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	0.5691	0.000	0.5197	0.6185	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.2317	0.000	-0.2811	-0.1823	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.3375	0.000	0.2881	0.3869	Yes
Groundedness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.1648	0.000	-0.1937	-0.1359	Yes
	No Refine vs MADR	No Refine	MADR	0.1689	0.000	0.1400	0.1978	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.1158	0.000	-0.1447	-0.0869	Yes
	No Refine vs SPP	No Refine	SPP	-0.0148	0.4793	-0.0437	0.0140	No
	MARA (Ours) vs MADR	MARA (Ours)	MADR	0.3337	0.000	0.3048	0.3626	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	0.2806	0.000	0.2517	0.3095	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.1796	0.000	-0.2085	-0.1508	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.1010	0.000	0.0721	0.1298	Yes
Naturalness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.1339	0.0001	-0.1909	-0.0770	Yes
	No Refine vs MADR	No Refine	MADR	0.3432	0.000	0.2863	0.4002	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.1511	0.000	-0.2081	-0.0942	Yes
	No Refine vs SPP	No Refine	SPP	-0.0277	0.5296	-0.0846	0.0293	No
	MARA (Ours) vs MADR	MARA (Ours)	MADR	0.4772	0.000	0.4202	0.5341	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	0.2851	0.000	0.2281	0.3420	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.1616	0.000	-0.2185	-0.1046	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.1235	0.0002	0.0665	0.1804	Yes
Engagingness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.6771	0.000	-0.7304	-0.6238	Yes
	No Refine vs MADR	No Refine	MADR	0.6168	0.000	0.5635	0.6701	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.1030	0.0006	-0.1563	-0.0497	Yes
	No Refine vs SPP	No Refine	SPP	0.3065	0.000	0.2532	0.3597	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	1.2939	0.000	1.2406	1.3472	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	0.7801	0.000	0.7268	0.8334	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.3706	0.000	-0.4239	-0.3173	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.4095	0.000	0.3562	0.4628	Yes

Table 9: Significance Testing for Coherence, Groundedness, Naturalness, and Engagingness (PersonaChat Dataset)

	Comparison	Model 1	Model 2	Mean Diff.	P-adj	Lower	Upper	Significant?
Coherence	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.4572	0.000	-0.5092	-0.4053	Yes
	No Refine vs MADR	No Refine	MADR	-0.2786	0.000	-0.3264	-0.2309	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	0.3962	0.000	0.3442	0.4481	Yes
	No Refine vs SPP	No Refine	SPP	0.1141	0.0002	0.0621	0.1660	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	0.7359	0.000	0.6839	0.7878	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	-0.0611	0.0204	-0.1130	-0.0091	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.3432	0.000	-0.3951	-0.2912	Yes
	Self-Refine vs SPP	Self-Refine	SPP	-0.2821	0.000	-0.3340	-0.2301	Yes
Groundedness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.1406	0.000	-0.1627	-0.1185	Yes
	No Refine vs MADR	No Refine	MADR	0.1325	0.000	0.1104	0.1546	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	0.1037	0.000	0.0816	0.1258	Yes
	No Refine vs SPP	No Refine	SPP	0.0215	0.0574	-0.0006	0.0436	No
	MARA (Ours) vs MADR	MARA (Ours)	MADR	0.2731	0.000	0.2510	0.2952	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	-0.0369	0.0019	-0.0590	-0.0148	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.1191	0.000	-0.1412	-0.0970	Yes
	Self-Refine vs SPP	Self-Refine	SPP	-0.0822	0.000	-0.1043	-0.0601	Yes
Naturalness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.4273	0.000	-0.4802	-0.3744	Yes
	No Refine vs MADR	No Refine	MADR	0.1728	0.000	0.1199	0.2257	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	0.5169	0.000	0.4640	0.5698	Yes
	No Refine vs SPP	No Refine	SPP	0.1673	0.000	0.1144	0.2202	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	0.6000	0.000	0.5471	0.6529	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	0.0896	0.0017	0.0367	0.1425	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.2599	0.000	-0.3128	-0.2070	Yes
	Self-Refine vs SPP	Self-Refine	SPP	-0.3496	0.000	-0.4025	-0.2967	Yes
Engagingness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	-0.9953	0.000	-1.0695	-0.9211	Yes
	No Refine vs MADR	No Refine	MADR	0.2260	0.000	0.1518	0.3002	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	0.7535	0.000	0.6793	0.8277	Yes
	No Refine vs SPP	No Refine	SPP	0.5208	0.000	0.4466	0.5950	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	1.2213	0.000	1.1471	1.2955	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	0.2418	0.000	0.1676	0.3160	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.4745	0.000	-0.5487	-0.4003	Yes
	Self-Refine vs SPP	Self-Refine	SPP	-0.2327	0.000	-0.3069	-0.1585	Yes

Table 10: Significance Testing for Coherence, Groundedness, Naturalness, and Engagingness (INSCIT Dataset)

	Comparison	Model 1	Model 2	Mean Diff.	P-adj	Lower	Upper	Significant?
Coherence	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	0.0844	0.0013	0.034	0.1348	Yes
	No Refine vs MADR	No Refine	MADR	-0.2420	0.000	-0.2925	-0.1916	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.2250	0.000	-0.2754	-0.1746	Yes
	No Refine vs SPP	No Refine	SPP	-0.0102	0.9529	-0.0582	0.0379	No
	MARA (Ours) vs MADR	MARA (Ours)	MADR	-0.3264	0.000	-0.3769	-0.2760	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	-0.3094	0.000	-0.3598	-0.2590	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.0946	0.0005	-0.1426	-0.0465	Yes
	MADR vs Self-Refine	MADR	Self-Refine	0.0171	0.7681	-0.0310	0.0651	No
	MADR vs SPP	MADR	SPP	0.2319	0.000	0.1839	0.2799	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.2148	0.000	0.1668	0.2629	Yes
Groundedness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	0.0765	0.000	0.0526	0.1005	Yes
	No Refine vs MADR	No Refine	MADR	-0.0525	0.0003	-0.0765	-0.0285	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.0721	0.0001	-0.0961	-0.0481	Yes
	No Refine vs SPP	No Refine	SPP	-0.0244	0.0518	-0.0491	0.0002	No
	MARA (Ours) vs MADR	MARA (Ours)	MADR	-0.1290	0.000	-0.1530	-0.1051	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	-0.1486	0.000	-0.1726	-0.1247	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.1010	0.000	-0.1256	-0.0764	Yes
	MADR vs Self-Refine	MADR	Self-Refine	-0.0196	0.1391	-0.0442	0.0050	No
	MADR vs SPP	MADR	SPP	0.0281	0.0245	0.0034	0.0527	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.0477	0.0006	0.0231	0.0723	Yes
Naturalness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	0.1052	0.0036	0.0365	0.1739	Yes
	No Refine vs MADR	No Refine	MADR	-0.2434	0.000	-0.3160	-0.1708	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.1727	0.0001	-0.2414	-0.1039	Yes
	No Refine vs SPP	No Refine	SPP	-0.0752	0.0309	-0.1439	-0.0065	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	-0.1382	0.0004	-0.2108	-0.0656	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	-0.0675	0.0549	-0.1362	0.0013	No
	MARA (Ours) vs SPP	MARA (Ours)	SPP	0.0300	0.6209	-0.0387	0.0987	No
	MADR vs Self-Refine	MADR	Self-Refine	0.0707	0.0430	-0.0018	0.1433	No
	MADR vs SPP	MADR	SPP	0.1682	0.0001	0.0995	0.2369	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.0975	0.0062	0.0287	0.1662	Yes
Engagingness	**No Refine vs MARA (Ours)**	No Refine	MARA (Ours)	0.4544	0.000	0.3586	0.5502	Yes
	No Refine vs MADR	No Refine	MADR	-0.3843	0.000	-0.4802	-0.2885	Yes
	No Refine vs Self-Refine	No Refine	Self-Refine	-0.0373	0.6615	-0.1331	0.0586	No
	No Refine vs SPP	No Refine	SPP	0.1880	0.0003	0.0981	0.2779	Yes
	MARA (Ours) vs MADR	MARA (Ours)	MADR	-0.8387	0.000	-0.9346	-0.7429	Yes
	MARA (Ours) vs Self-Refine	MARA (Ours)	Self-Refine	-0.4917	0.000	-0.5875	-0.3958	Yes
	MARA (Ours) vs SPP	MARA (Ours)	SPP	-0.2664	0.000	-0.3563	-0.1765	Yes
	MADR vs Self-Refine	MADR	Self-Refine	0.3471	0.000	0.2512	0.4429	Yes
	MADR vs SPP	MADR	SPP	0.5723	0.000	0.4824	0.6623	Yes
	Self-Refine vs SPP	Self-Refine	SPP	0.2253	0.0001	0.1354	0.3152	Yes

B Experimental Results

B.1 Significance Test on FoCus (Table 8)

Coherence The one-way ANOVA reveals a significant effect of model type on coherence scores ($F(3, 8) = 763.67, p < 0.001$), indicating that at least one model has a significantly different mean score. Post-hoc Tukey HSD tests show that MARA significantly outperforms all other models ($p < 0.001$), with a mean difference of 0.7456 over the MADR model, 0.2838 over the No Refine model, and 0.5691 over the Self-Refine model. Additionally, the No Refine model outperforms both Self-Refine and SPP, with a small but significant difference of 0.0521 over SPP. Self-Refine also outperforms SPP with a mean difference of 0.3375. Overall, MARA is the top performer, offering substantial improvements in coherence, demonstrating that its enhancements contribute meaningfully to better response quality across all models.

Groundedness The one-way ANOVA reveals a significant effect of model type on groundedness scores ($F(3, 8) = 426.81, p < 0.001$), indicating notable differences in performance among the models. Post-hoc Tukey HSD tests show that MARA achieves higher groundedness scores than all other models ($p < 0.001$), with mean differences of 0.3337 over MADR, 0.1648 over No Refine, and 0.2806 over Self-Refine. Additionally, No Refine surpasses Self-Refine by 0.1158, while its difference from SPP is not significant. Self-Refine also shows an advantage over SPP, with a mean difference of 0.101. These results highlight the effectiveness of MARA in improving groundedness compared to other approaches.

Naturalness The one-way ANOVA reveals a significant effect of model type on naturalness scores ($F(3, 8) = 215.79, p < 0.001$), indicating that model differences lead to varying levels of naturalness. Post-hoc Tukey HSD tests show that

MARA significantly outperforms all other models ($p < 0.001$), with mean differences of 0.4772 over MADR, 0.1339 over No Refine, and 0.2851 over Self-Refine. Furthermore, No Refine performs better than Self-Refine by 0.1511, although its comparison with SPP shows no significant difference. Self-Refine outperforms SPP with a mean difference of 0.1235. Overall, these results indicate that MARA demonstrates strong naturalness compared to other models, with clear advantages over the other approaches.

Engagingness The one-way ANOVA reveals a significant effect of model type on engagingness scores ($F(3, 8) = 1772.37, p < 0.001$), indicating notable differences across the models. Post-hoc Tukey HSD tests show that MARA significantly outperforms all other models ($p < 0.001$), with mean differences of 1.2939 over MADR, 0.6771 over No Refine, and 0.7801 over Self-Refine. Furthermore, No Refine outperforms Self-Refine by 0.1030, and SPP by 0.3065. Self-Refine also outperforms SPP with a mean difference of 0.4095. Overall, these results demonstrate that MARA consistently achieves higher engagingness scores compared to the other models, highlighting its effectiveness in enhancing the engagingness of responses.

B.2 Significance Test on PersonaChat (Table 9)

Coherence The one-way ANOVA reveals a significant effect of model type on coherence scores ($F(3, 8) = 725.87, p < 0.001$), indicating substantial differences between the models. Post-hoc Tukey HSD tests show that MARA significantly outperforms all other models ($p < 0.001$), with mean differences of 0.7359 over MADR, 0.4572 over No Refine, and 0.0611 over Self-Refine. Additionally, No Refine outperforms Self-Refine by 0.3962 and SPP by 0.1141. Self-Refine also outperforms SPP with a mean difference of 0.2821. These results demonstrate that MARA leads to higher coherence scores, providing a clear advantage over the other models in this aspect.

Groundedness The one-way ANOVA reveals a significant effect of model type on groundedness scores ($F(3, 8) = 498.84, p < 0.001$), indicating notable differences between the models. Post-hoc Tukey HSD tests show that MARA significantly outperforms all other models ($p < 0.001$), with mean differences of 0.2731 over MADR, 0.1406 over No Refine, and 0.0369 over Self-Refine. Furthermore, No Refine outperforms Self-Refine by 0.1037, although its comparison with SPP shows no significant difference. Self-Refine also outperforms SPP with a mean difference of 0.0822. These results indicate that MARA provides consistent improvements in groundedness compared to other models.

Naturalness The one-way ANOVA reveals a significant effect of model type on naturalness scores ($F(3, 8) = 641.13, p < 0.001$), indicating substantial differences across the models. Post-hoc Tukey HSD tests show that MARA significantly outperforms all other models ($p < 0.001$), with mean differences of 0.6000 over MADR, 0.4273 over No Refine, and 0.2599 over SPP. The comparison between MARA and Self-Refine also shows a small but significant difference of 0.0896. Furthermore, No Refine outperforms Self-Refine by 0.5169 and SPP by 0.1673, while Self-Refine significantly outperforms SPP by 0.3496. These results highlight MARA's strong performance in naturalness compared to other models.

Engagingness The one-way ANOVA reveals a significant effect of model type on engagingness scores ($F(3, 8) = 1028.47, p < 0.001$), showing that the models differ significantly in their performance. Post-hoc Tukey HSD tests indicate that MARA significantly outperforms all other models ($p < 0.001$), with mean differences of 1.2213 over MADR, 0.9953 over No Refine, and 0.4745 over SPP. Additionally, No Refine outperforms Self-Refine by 0.7535 and SPP by 0.5208, while Self-Refine outperforms SPP by 0.2327. These results underscore MARA's effectiveness in enhancing engagingness across different models.

B.3 Significance Test on INSCIT (Table 10)

Coherence The one-way ANOVA reveals a significant effect of model type on coherence scores ($F(3, 8) = 200.85, p < 0.001$), indicating that the models perform differently. Post-hoc Tukey HSD tests show that MARA significantly outperforms MADR, with a mean difference of 0.3264, as well as No Refine (0.0844) and SPP (0.0946). The comparison between MADR and Self-Refine is not statistically significant, but No Refine outperforms Self-Refine by 0.225. Self-Refine also outperforms SPP with a mean difference of 0.2148. Overall, MARA demonstrates stronger coherence performance compared to most other models.

Table 11: Case study on the FoCus dataset comparing MARA with the second-best performing refinement model, SPP. Misalignments are highlighted in red, and alignments in blue.

Question	Persona	Fact	SPP	MARA (Ours)
Q1: What is the height of the mountain?	I have the fantasy about valley. I like lakes. I would like to go to California. I have seen lava domes. I have heard about Lassen Volcanic National Park.	Located 2 miles (3.2 km) to the north of Lassen Peak, they have an elevation of about 8,448 feet (2,575 m)	Unfortunately, without any additional context about what fictional universe, book, movie or game Chaos Crags is from, it's very difficult for me to estimate the height of this mountain. If you could provide some more details about the setting, terrain, or role this mountain plays in the story, that would help me make an educated guess about how tall Chaos Crags is likely intended to be. I'd be happy to take another look once I have some contextual clues from the source material.	The Chaos Crags are not a single mountain, but rather a group of six steep-sided lava domes or volcanic plugs. They rise approximately 550 meters (about 1,800 feet) above the Chaos Jumbles, which is the rockfall avalanche deposit at their base. To provide more context, the highest point of the Chaos Crags reaches an elevation of about 2,575 meters (8,450 feet) above sea level.
Q2: Is there any lake near by?	I have the fantasy about valley. I like lakes. I would like to go to California. I have seen lava domes. I have heard about Lassen Volcanic National Park.	Each year, a lake forms at the base of the Crags, and typically dries by the end of the summer season.	Based on the lore provided, Chaos Crags is an immense, treacherous mountain range in the northern reaches of Middle-earth. While the exact height is not specified, it is described as one of the tallest peaks in that region, with the highest spires potentially reaching elevations of 20,000-25,000 feet. The Crags likely consist of jagged, ice-capped pinnacles towering miles into the sky above deep glacial valleys. Their sheer cliffs and avalanche-prone slopes would make them extremely hazardous and virtually unscalable. While no major lakes are referenced, the Crags overlook the icy waters of the Bay of Forochel and its inlets. Nestled among the rocky chaos may be smaller glacier-fed tarns and streams cascading from the ice fields high above. The overall scenery is one of bleak, desolate beauty - an otherworldly alpine expanse exposed to the punishing winds blowing off the northern seas.	As someone drawn to the beauty of lakes and the raw power of volcanic landscapes, you'll be delighted to know there are several lakes nestled right near the incredible Chaos Crags formation in Lassen Volcanic National Park. The most picturesque is undoubtedly Manzanita Lake, formed around 1,100 years ago by lava flows that dammed a stream during the same eruptive period that created the crags themselves. Can you imagine hiking along its shores, taking in views of those dramatic lava domes rising in the backdrop? A scene like that perfectly encapsulates the awe-inspiring volcanic wonders of this region. But Manzanita isn't the only lake beckoning explorers like yourself. Reflection Lake and Chaos Crags Lake, though smaller, offer equally enticing opportunities to immerse yourself in this geologically active landscape.

Groundedness The one-way ANOVA reveals a significant effect of model type on groundedness scores ($F(3, 8) = 119.45, p < 0.001$), indicating that the models show distinct performance differences. Post-hoc Tukey HSD tests show that MARA significantly outperforms MADR with a mean difference of 0.129, as well as No Refine (0.0765) and SPP (0.101). No Refine also outperforms Self-Refine by 0.0721. However, the comparison between MADR and Self-Refine is not statistically significant. These results suggest that MARA provides meaningful improvements in groundedness compared to the other models.

Naturalness The one-way ANOVA reveals a significant effect of model type on naturalness scores ($F(3, 8) = 39.70, p < 0.001$), indicating differences in performance across models. Post-hoc Tukey HSD tests show that MARA significantly outperforms MADR by 0.1382 and No Refine by 0.1052, though the comparison between MARA and Self-Refine is not significant. MADR also outperforms No Refine by 0.2434 and Self-Refine by 0.0707, while No Refine outperforms Self-Refine by 0.1727. These results demonstrate that MARA provides strong naturalness performance, with notable differences in certain comparisons, but not against all models.

Table 12: Analyses in the FoCus dataset.

	Coh.	Grd.	Nat.	Eng.
MARA (Ours)	**2.67**	**0.65**	**2.15**	**2.83**
w/o Planner Outputs	2.62	0.60	2.12	2.75
w/o Step-by-Step	2.59	0.65	2.00	2.66
No Refine	2.39	0.49	2.02	2.15

Engagingness The one-way ANOVA reveals a significant effect of model type on engagingness scores ($F(3, 8) = 255.26, p < 0.001$), showing differences in model performance. Post-hoc Tukey HSD tests show that MARA significantly outperforms MADR by 0.8387, No Refine by 0.4544, and SPP by 0.2664. The comparison between No Refine and Self-Refine is not statistically significant, but No Refine outperforms SPP by 0.188, and Self-Refine outperforms SPP by 0.2253. These results suggest that MARA provides stronger engagingness performance compared to the other models, with significant improvements over most.

B.4 Ablation Studies on Design Choices

To further investigate the effectiveness of our strategy, we conduct ablation studies on specific design choices. First, we examine the impact of allowing each refining agent to recognize the presence of other agents by passing the planner agent's output, which includes both the sequence and justification for the selected set and sequence. Specifically, the planner agent generates a suitable sequence with justification and passes it to the refining agents, enabling each agent to be aware of the preceding and following agents, as well as the rationale behind the sequence. As shown in Table 12, when the planner agent's output is not passed to the refining agents, performances with all metrics decrease, particularly in groundedness. This underscores the importance of enabling agents to be aware of one another to perform their roles and collaborate more effectively.

Additionally, we explore whether a step-by-step for each refining agent's reasoning process—first verifying, then refining responses with justifications—is effective. As shown in Table 12, performance in coherence, naturalness, and engagingness decrease when this process is removed. This suggests that the verification step and justification is crucial for maintaining high-quality responses, as it allows for a more thorough assessment before refining.

B.5 Case Study

We conduct a case study to qualitatively analyze the effectiveness of our proposed method. Table 11 shows two consecutive turns within a conversation about the 'Chaos Crags'. In the first turn, SPP fails to provide an accurate response, losing track of the conversation's context and offering irrelevant information about fictional mountains instead of directly addressing the user's query about the mountain's height. In contrast, our MARA model delivers an accurate and contextually appropriate response, correctly identifying Chaos Crags and providing the exact elevation, enriched with additional context about its geological formation. This demonstrates MARA's ability to maintain focus and generate factually grounded responses.

In the second turn, both SPP and MARA attempt to understand the context of the user's question. However, SPP not only fails to directly address the user's question about a nearby lake, stating, 'While no major lakes are referenced,' but also lacks engagement with the user. MARA, on the other hand, not only provides the correct factual information but also tailors the response to the user's expressed interest in lakes and volcanic landscapes, creating a more personalized and engaging interaction. Furthermore, MARA takes the conversation a step further by asking a follow-up question: 'Can you imagine hiking along its shores, taking in views of those dramatic lava domes rising in the backdrop?' This question invites the user to visualize the experience, fostering deeper engagement. Such interactions showcase MARA's strength in incorporating user preferences and enhancing conversational quality beyond simple fact delivery.

Table 13: The prompt used in the full instantiation of MARA for responding agent.

Types	Texts
System Message	As a <role>Responding Agent</role>, your task is to answer the user's question, within the <question_text> tags. - Consider the Keyword: <keyword>{keyword}</keyword>, if available. <instructions> - Place your final response within <response></response> tags. - Make your response concise. </instructions>
User Message	<question_text>{user_query}</question_text>

Table 14: The prompt used in the full instantiation of MARA for the planner agent.

Types	Texts
System Message	Role: <role>Planner Agent</role>. Your task is to select the appropriate Agent(s) to refine the response to the user's query, following a step-by-step reasoning process. <considerations> - <userProfile>{persona}</userProfile>: Understand the user's profile to tailor your response. - <keywords>{keyword}</keywords>: Identify the key topics and context of the conversation. </considerations> <instructions> - Begin by examining the user's interests and the conversation's context as detailed in the <userProfile> and <keywords> tags. Also review the user's query in the <question_text> tags and the initial response in the <initialResponse> tags. - Next, read the roles of the three Refining Agents available, each tasked with specific aspects of response refinement: <Persona Refining Agent> Verifies and refines the alignment with the user's profile and interests. </Persona Refining Agent> <Coherence Refining Agent> Verifies and refines the coherence of responses. </Coherence Refining Agent> <Fact Refining Agent> Verifies and refines the factual accuracy of responses. </Fact Refining Agent> - Ensure that the response to the user's query in the <question_text> tags is factually accurate, aligns with the user's profile, and is coherent. - Most importantly, ensure the response fully addresses the user's question inside the <question_text> tags. - Based on your analysis of each Agent's contributions, decide which Agent or combination of Agents is best suited to refine the response in the <agents_set> tag. - Determine the optimal sequence of Agent involvement based on the necessary refinements. Write the order in <agents_set>, separating each agent with a comma character (', '). Articulate your reasoning in the <agents_set_justification> tags. - Also note that the sequence should reflect the priorities in refining the response to make it as relevant and accurate as possible. Also include the justification of the order in <agents_set_order_justification>. - If the initial response is sufficient and no further refinement is needed, write 'None' in the <agents_set> tag. - Place <agents_set>, <agents_set_order_justification>, and <agents_set_justification> as your final response in <agent_planning> tags. </instructions>
User Message	<question_text>{user_query}</question_text> <initialResponse>{initial_response}</initialResponse>

Table 15: The prompt used in the full instantiation of MARA for fact-refining agent.

Types	Texts
System Message	Role: <role>Fact Refining Agent</role>. Your task is to refine the previous responses to ensure they are accurate within the context of the conversation, following a step-by-step reasoning process. Consider the given topics (keywords) for the conversation, if available, marked by <keywords> tags: <keywords>{keyword}</keywords>. <instructions> - Begin by reviewing the keywords in the <keywords> tags to understand the context of the conversation. Note that we assume that the user's opening question specifically focuses on the keywords listed within the <keywords> tags. Therefore, do not request additional detail or clarification. - First, examine the factual accuracy of the previous response provided in the <factChecking> tags. - Document your verification outcome: If the response is factually accurate, place 'Fact is verified.' inside the <verification> tags. If not, place 'Fact is not verified.' in the <verification> tags. - Specify the reasons for the verification in the <verification_justification> tags. - Next, refine the previous response to enhance factual correctness in the <refined_response> tags. Describe each change you make and justify it based on factual accuracy. - Place the refined response in the <refined_response> tags and detail your reasoning for each refinement step in <refinement_justification> tags. - Finally, compile the <verification>, <verification_justification>, <refined_response> and <refinement_justification> into the <response> tags, ensuring a clear and logical flow of thought. - Maintain professionalism and avoid including apologies or acknowledgements in your response. - Since this refined response is displayed directly to the user, address them as if you are the Responding Agent, not a Refining Agent. Avoid any reference to refining roles, the refinement process, or acknowledgment of a previous Refining Agent. - Ensure your response shows you have considered the user's input inside the <question_text> tags and their current state of mind. Keep your <refined_response> concise, clear, and similar to human-written text. </instructions>
User Message	This is the sequence of multiple agents involved in refining the initial response: {planned_agent_order}. This is the justification for requiring multiple agents: {planned_agents_set_justification}. This is the justification for this sequence: {planned_agent_order_justification}. Note that you are a Fact Refining Agent. This is the user's question, which should be fully addressed: <question_text>{user_query}</question_text> This is the initial response generated by the Responding Agent: <initialResponse>{initial_response}</initialResponse>. This is the refined response generated by the previous refining agent ({previous_agent_name}): {generated_response}.

Table 16: The prompt used in the full instantiation of MARA for persona-refining agent.

Types	Texts
System Message	Role: <role>Persona Refining Agent</role>. Your task involves verifying and refining responses based on the user's profile and interests. You will first verify the alignment and then refine the response if necessary, explaining each step in your reasoning. Consider the User Profile: <userProfile>{persona}</userProfile> <instructions> - Begin by understanding the user's interests from the <userProfile> tags. - First, verify the alignment of the previous response inside the <persona> tags aligns with the user's preferences detailed in the <userProfile> tags. Explain your reasoning: If it aligns well, explain why you believe this alignment exists; if it does not, explain what aspects are misaligned or contradictory. - Document your verification outcome: If the previous response is well-aligned, place 'Persona is verified.' inside the <verification> tags. If not, place 'Persona is not verified.' inside the <verification> tags. - Specify the reasons for the verification in the <verification_justification> tags. - Next, refine the previous response to better align with the user's preferences detailed in the <userProfile> tags. - Place the refined response in the <refined_response> tags and detail your reasoning for each refinement step in the <refinement_justification> tags. - Finally, compile the <verification>, <verification_justification>, <refined_response> and <refinement_justification> into the <response> tags, ensuring a clear and logical flow of thought. - Maintain professionalism and avoid including apologies or acknowledgements in your response. - Since this refined response is displayed directly to the user, address them as if you are the Responding Agent, not a Refining Agent. Avoid any reference to refining roles, the refinement process, or acknowledgment of a previous Refining Agent. - Ensure your response shows you have considered the user's input inside the <question_text> tags and their current state of mind. Keep your <refined_response> concise, clear, and similar to human-written text. </instructions>
User Message	This is the sequence of multiple agents involved in refining the initial response: {planned_agent_order}. This is the justification for requiring multiple agents: {planned_agents_set_justification}. This is the justification for this sequence: {planned_agent_order_justification}. Note that you are a Persona Refining Agent. This is the user's question, which should be fully addressed: <question_text>{user_query}</question_text> This is the initial response generated by the Responding Agent: <initialResponse>{initial_response}</initialResponse>. This is the refined response generated by the previous refining agent ({previous_agent_name}): {generated_response}.

Table 17: The prompt used in the full instantiation of MARA for coherence-refining agent.

Types	Texts
System Message	Role: <role>Coherence Refining Agent</role>. Your task is to verify and refine the previous responses to ensure the coherence and logical flow of the initial responses within the context of the ongoing conversation and mimicking human-written text, following a step-by-step reasoning process. Consider the given topics (keywords) for the conversation, if available, marked by <keywords> tags: <keywords>{keyword}</keywords>. <instructions> - Begin by reviewing the keywords in the <keywords> tags to understand the context of the conversation. Note that we assume that the user's opening question specifically focuses on the keywords listed within the <keywords> tags. Therefore, do not request additional detail or clarification. - First, examine whether the previous response provided in the <coherence> tags maintains coherence with the conversation history, mimicking human-written text. - Document your verification outcome: If the response is coherent, place 'Coherence is verified.' inside the <verification> tags. If not, place 'Coherence is not verified.' in the <verification> tags. - Specify the reasons for the verification in the <verification_justification> tags. - Next, refine the previous response to improve the overall clarity and continuity of the conversation in the <refined_response> tags. Describe each change you make and justify it based on coherence and completeness. - Place the refined response in the <refined_response> tags and detail your reasoning for each refinement step in <refinement_justification> tags. - Finally, compile the <verification>, <verification_justification>, <refined_response> and <refinement_justification> into the <response> tags, ensuring a clear and logical flow of thought. - Maintain professionalism and avoid including apologies or acknowledgements in your response. - Since this refined response is displayed directly to the user, address them as if you are the Responding Agent, not a Refining Agent. Avoid any reference to refining roles, the refinement process, or acknowledgment of a previous Refining Agent. - Ensure your response shows you have considered the user's input inside the <question_text> tags and their current state of mind. Keep your <refined_response> concise, clear, and similar to human-written text. </instructions>
User Message	This is the sequence of multiple agents involved in refining the initial response: {planned_agent_order}. This is the justification for requiring multiple agents: {planned_agents_set_justification}. This is the justification for this sequence: {planned_agent_order_justification}. Note that you are a Coherence Refining Agent. This is the user's question, which should be fully addressed: <question_text>{user_query}</question_text> This is the initial response generated by the Responding Agent: <initialResponse>{initial_response}</initialResponse>. This is the refined response generated by the previous refining agent ({previous_agent_name}): {generated_response}.

Table 18: The prompt template used for G-Eval (Coherence).

Instructions:

You will be given a conversation segment involving two participants: the user and the system. You will then be given one potential response for the next turn in the conversation. Your task is to rate the generated responses on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-3) - Evaluate whether the conversation response logically follows the preceding context and maintains a clear, logical flow.

Score of 1: Assign this score if the response does not logically follow the preceding context. It may introduce abrupt changes in topic or contain confusing statements, resulting in a disjointed conversation.

Score of 2: Assign this score if the response somewhat follows the preceding context but includes minor logical inconsistencies or slight topic shifts, slightly disrupting the conversation's flow.

Score of 3: Assign this score if the response logically follows the preceding context and maintains a clear, logical flow, providing a relevant and coherent continuation of the conversation without any abrupt changes or confusing elements.

Evaluation Steps:

1. Examine the conversational history, the provided fact (if given), and the user profile (if given) to fully understand the context and dynamics of the conversation.

2. Compare the generated response with the gold standard response to evaluate how well it maintains continuity and logical flow.

3. Assess how effectively the generated response connects with and continues the conversation, ensuring it aligns logically with the existing conversation.

4. Based on your analysis, assign a coherence score from 1 to 3, reflecting the response's logical integration into the ongoing conversation.

Example:

Conversation History: {{Document}}

Corresponding Fact: {{Fact}}

Corresponding User Profile: {{Persona}}

Gold Standard Response: {{Gold_Response}}

Generated Response: {{Response}}

Evaluation Form (Scores ONLY without any additional text):

- Coherence:

Table 19: The prompt template used for G-Eval (Groundedness).

Instructions:

You will be given a conversation segment involving two participants: the user and the system. You will then be given one potential response for the next turn in the conversation. The response concerns an interesting fact, which will be provided as well. Your task is to rate the generated responses on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Groundedness (0-1) - Evaluate whether the conversation response is based on and accurately incorporates the provided fact.

Score of 0: Assign this score if the response does not correctly use the provided fact or misrepresents it. This includes instances where the response contains inaccuracies or fails to integrate the fact meaningfully into the conversation.

Score of 1: Assign this score if the response uses the provided fact accurately and integrates it seamlessly into the conversation, thereby enhancing the dialogue's relevance and informativeness.

Evaluation Steps:

1. Examine the conversational history and the provided fact (if given) to understand the context fully.

2. Evaluate how accurately and relevantly the generated response incorporates the provided fact, ensuring there are no inaccuracies or hallucinated details.

3. Analyze how the generated response measures up against a gold standard response to understand the ideal integration of the fact.

4. Based on your assessment, assign a Groundedness score ranging from 0 to 1, reflecting how effectively the fact is incorporated into the response.

Example:

Conversation History: {{Document}}

Corresponding Fact: {{Fact}}

Gold Standard Response: {{Gold_Response}}

Generated Response: {{Response}}

Evaluation Form (Scores ONLY without any additional text):

- Groundedness:

Table 20: The prompt template used for G-Eval (Naturalness).

Instructions

You will be given a conversation segment involving two participants: the user and the system. You will then be given one potential response for the next turn in the conversation. Your task is to rate the generated responses on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Naturalness (1-3) - Evaluate whether the dialogue response feels natural and conversational, as if it were part of a real, human conversation.

Score of 1: Assign this score if the response does not sound natural. It may contain awkward phrasing, unnatural expressions, or robotic language, disrupting the flow of the conversation.

Score of 2: Assign this score if the response somewhat sounds natural but may include minor awkwardness or slightly unnatural phrasing, affecting the overall conversational flow.

Score of 3: Assign this score if the response sounds completely natural, flowing smoothly, using natural language, and integrating seamlessly into the conversation as if it were part of a real human interaction.

Evaluation Steps:

1. Examine the conversational history, the provided fact (if given), and the user profile (if given) to gauge the natural fit of the response within the conversation's context.

2. Evaluate the tone, formality, and conversational flow of the generated response to determine how naturally it fits into the dialogue.

3. Compare the generated response to a gold standard response to gauge the ideal level of naturalness.

4. Based on your assessment, assign a Naturalness score from 1 to 3, focusing on how naturally the response fits into the conversation.

Example:

Conversation History: {{Document}}

Corresponding Fact: {{Fact}}

Corresponding User Profile: {{Persona}}

Gold Standard Response: {{Gold_Response}}

Generated Response: {{Response}}

Evaluation Form (Scores ONLY without any additional text):

- Naturalness:

Table 21: The prompt template used for G-Eval (Engagingness).

Instructions

You will be given a conversation segment involving two participants: the user and the system. You will then be given one potential response for the next turn in the conversation. Your task is to rate the generated responses on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Engagingness (1-3) - Is the response dull or interesting?

Score of 1 (Dull): Assign this score if the response is generic and unremarkable, failing to spark interest or engagement.

Score of 2 (Somewhat Interesting): Assign this score if the response is moderately interesting and could engage participants in the conversation, such as by introducing an opinion or thought.

Score of 3 (Interesting): Assign this score if the response is highly interesting or presents an intriguing fact, significantly enhancing the conversation's appeal.

Evaluation Steps:

1. Examine the conversational history, the provided fact (if given), and the user profile (if given) to gauge the potential interest or intrigue.

2. Assess how the generated response contributes to the conversation's value and captivates interest.

3. Compare the generated response to a gold standard response.

4. Based on your analysis, assign an Engagingness score from 1 to 3, reflecting the response's ability to captivate and add value to the conversation.

Example:

Conversation History: {{Document}}

Corresponding Fact: {{Fact}}

Corresponding User Profile: {{Persona}}

Gold Standard Response: {{Gold_Response}}

Generated Response: {{Response}}

Evaluation Form (Scores ONLY without any additional text):

- Engagingness: