CEUR-WS.org/Vol-4178/paper3.pdf

Adaptive Multi-Agent Response Refinement in Conversational Systems

Soyeong Jeong[1,†], Aparna Elangovan[3,‡], Emine Yilmaz[2,4] and Oleg Rokhlenko[2]

[1]KAIST
[2]Amazon
[3]Collate
[4]University College London

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 personalization 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 diverse aspects needed for effective conversations. In this work, we propose refining responses through a multi-agent framework, where each agent is assigned a specific role for each aspect. We focus on three key aspects crucial to conversational quality: factuality, personalization, 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 them, 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 challenging conversational datasets, demonstrating that ours significantly outperforms relevant baselines, particularly in tasks involving knowledge or user's persona, or both.

Keywords

Conversational AI, Multi-Agent Systems, Adaptive Response Refinement, Large Language Models

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 [1, 2]. However, LLMs do not always produce satisfactory responses on the first attempt [3, 4], 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 [5, 6, 7]. These difficulties are particularly pronounced in complex, personalized conversations that demand both user's persona alignment and factual accuracy [8, 9]. 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 [10, 11, 12]. 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 [10], where a single agent handles an entire refinement process, generating feedback across multiple aspects and iteratively improving its outputs based on it. However, relying solely on a single agent can be suboptimal, as the agent may become overly confident in the initial output, leading to bias [13], which can be amplified through repeated iterations [14, 15]. This issue could be particularly critical for multi-turn conversational tasks, where errors in earlier turns can propagate to subsequent turns [16].

LaCATODA 2026: The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop at the 40th AAAI conference

[†] Work done during internship at Amazon.

[‡] Work done while at Amazon.

starsuzi@kaist.ac.kr (S. Jeong)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-03-04

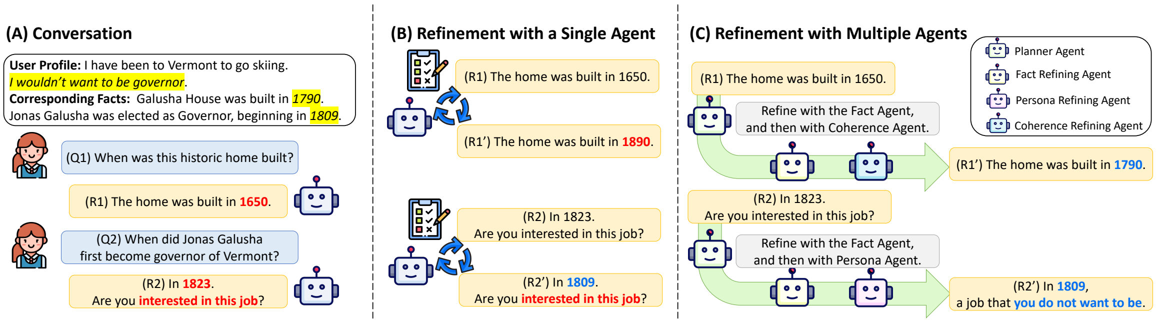
(A) Conversation
User Profile: I have been to Vermont to go skiing.
I wouldn't want to be governor.
Corresponding Facts: Galusha House was built in 1790.
Jonas Galusha was elected as Governor, beginning in 1809.
(Q1) When was this historic home built?
(R1) The home was built in 1650.
(Q2) When did Jonas Galusha first become governor of Vermont?
(R2) In 1823. Are you interested in this job?
(B) Refinement with a Single Agent
(R1) The home was built in 1650.
(R1') The home was built in 1890.
(R2) In 1823. Are you interested in this job?
(R2') In 1809. Are you interested in this job?
(C) Refinement with Multiple Agents
Planner Agent
Fact Refining Agent
Persona Refining Agent
Coherence Refining Agent
(R1) The home was built in 1650.
Refine with the Fact Agent, and then with Coherence Agent.
(R1') The home was built in 1790.
(R2) In 1823. Are you interested in this job?
Refine with the Fact Agent, and then with Persona Agent.
(R2') In 1809, a job that you do not want to be.

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.

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 sub-tasks and enabling collaborative problem-solving [17, 18, 19]. Additionally, the multi-agent framework enables for integration of a broader range of tools and knowledge sources [20, 21, 22], ensuring that each agent can be specifically tailored to meet the diverse needs of each task. While multi-agent frameworks have shown 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 a response should accurately reflect the user profile and specific knowledge. In such setting, employing multiple refining agents with specialized perspectives can 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: user 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 challenging conversational datasets that involve user persona integration, require specific factual knowledge, or both. The results show that MARA significantly outperforms relevant baselines, indicating that refining along the three aspects, persona, factuality, and coherence, is beneficial. Moreover, our analyses highlight the importance of dynamically selecting appropriate refinement strategies tailored to the specific needs of each query in diverse conversational contexts.

2. Related Work

LLMs in Conversation Systems. Recent Large Language Models (LLMs) have demonstrated remarkable performance across a range of NLP tasks [23, 24, 25], even without additional training, largely due to their massive pretraining on diverse datasets [26, 27]. Their success is also remarkable in conversational tasks [28, 1, 2], as LLMs are able to generate human-like responses [29, 30, 11]. However, LLMs occasionally produce unsatisfactory responses, particularly when they fail to consider a user's profile [9], specific knowledge [8], or are not coherent with the previous turns [31, 32].

Verification and Refinement with LLMs. To address the issue of unsatisfactory responses, several studies have explored verification and refinement approaches. While some research involves the use of additional refinement models, [33, 34], more recent studies emphasize the capability of LLMs to verify and refine their own responses without the need for further training [35, 36, 37, 38, 39]. Specifically, Madaan et al. [10] proposed a method where a single agent LLM generates feedback on multiple aspects of its own response. However, single-agent-based refinements can be less effective, as the overall performance is constrained by the capacity of a single agent, and once an agent becomes confident in its outputs, it may struggle to further generate novel thoughts [13, 14, 15].

Multi-agent LLMs. Along with the powerful 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 [40, 17, 18, 41]. Multi-agent LLM framework has recently been widely applied to various tasks, including software development [42, 43], model evaluation [44], research topic generation [45], diagnostic consultations [46, 47], recommendation [48, 49], and reasoning improvement [50]. However, the use of multiple agents for refining conversational turns remains underexplored.

3. Method

We introduce MARA, a multi-agent framework that dynamically refines conversational responses.

3.1. Preliminaries

We begin by formally defining an LLM, specifically in the context of a multi-turn conversational setting.

Large Language Models. Let us first define LLM as `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 $\boldsymbol{y} = \texttt{LLM}(\mathscr{P}(\boldsymbol{x}))$, where the prompt template $\mathscr{P}$ incorporates additional context or instructions that guide the LLM's behavior. Specifically, $\mathscr{P}$ can be used for role assignment, ensuring that the LLM adopts a particular role or performs a specific task.

Conversation with an LLMs. 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: $\boldsymbol{r}^i = \texttt{LLM}(\mathscr{P}_{\text{respond}}(\boldsymbol{q}^i, \boldsymbol{q}^{i-1}, \boldsymbol{r}^{i-1}, ..., \boldsymbol{q}^1, \boldsymbol{r}^1))$, where $\boldsymbol{q}^i$ is the user's query at the i-th turn, and $\boldsymbol{r}^i$ is the response generated by the `LLM` which operates under the prompt template $\mathscr{P}_{\text{respond}}$. However, the quality of initial responses from the LLM may be suboptimal in customized, realistic conversations, potentially overlooking the user's profile, missing specific knowledge, or failing to maintain coherence, and such errors in earlier conversational turns can cumulatively 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.

Single-Agent Response Refinement. As an initial response $\boldsymbol{r}$ may be inaccurate or unsatisfactory, the goal of response refinement is to further improve its quality. This can be achieved by using an LLM operating as a refining agent, represented as: $\boldsymbol{r}_{\texttt{refine}} = \texttt{LLM}(\mathscr{P}_{\texttt{refine}}(\boldsymbol{r}))$, where an LLM instantiated with $\mathscr{P}_{\texttt{refine}}$ refines an initial response $\boldsymbol{r}$ to produce the improved response $\boldsymbol{r}_{\texttt{refine}}$. However, relying on a single agent to assess and refine multiple aspects can be ineffective, as it may struggle to holistically address the diverse factors essential for a high-quality, customized conversation. To overcome these limitations, we propose a multi-agent framework in which specialized refining agents collaborate to enhance response quality.

Multiple Refining Agents for Conversation. In a realistic yet challenging conversational scenario, responses must be contextually relevant, aligned with user preferences, and factually reliable. To meet these requirements, responses must be refined across multiple dimensions, ensuring factual correctness, alignment with the user's persona, and coherence across multiple conversational turns. Here, instead of addressing all these aspects with a single agent, we introduce three specialized refining agents that extend the general refining agent, $\texttt{LLM}(\mathscr{P}_{\texttt{refine}}(\boldsymbol{r}))$, each addressing a distinct aspect of response quality. Specifically, the fact-refining agent generates the refined response, $\boldsymbol{r}_{\texttt{fact-refine}}$, that ensures factual accuracy, $\boldsymbol{r}_{\texttt{persona-refine}}$ that aligns responses with the user's profile, and $\boldsymbol{r}_{\texttt{coherence-refine}}$ that maintains coherence throughout conversation, where each $\texttt{LLM}$ is instantiated with its respective refining role template: $\mathscr{P}_{\texttt{fact-refine}}$, $\mathscr{P}_{\texttt{persona-refine}}$, and $\mathscr{P}_{\texttt{coherence-refine}}$. Having defined the role of each refining agent, the next key challenge lies in how these agents should collaborate to collectively refine the response.

3.3. Communication Strategy among Agents

To enable effective collaboration among multiple agents, we explore various communication strategies and propose a dynamic approach to optimize the refinement process.

Simultaneous Communication. As an initial approach to agent collaboration, we introduce a simultaneous communication strategy, where all refining agents independently refine the initial response, and their refined outputs are then passed to a finalizer agent instantiated with $\mathscr{P}_{\texttt{finalizer}}$, which aggregates the individual refinements into a single unified response. Formally, this process is denoted as: $\texttt{LLM}(\mathscr{P}_{\texttt{finalizer}}(\boldsymbol{r}_{\texttt{fact-refine}}, \boldsymbol{r}_{\texttt{persona-refine}}, \boldsymbol{r}_{\texttt{coherence-refine}}))$. 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 $\boldsymbol{r}$, it is sequentially refined by a series of agents denoted as $\boldsymbol{s} = [\texttt{LLM}(\mathscr{P}_{\texttt{refine}_k}(\boldsymbol{r}))]_{k=1}^{n}$, where each refining agent is instantiated with a role-specific prompt template $\mathscr{P}_{\texttt{refine}_k}$, selected from the set of refinement templates: $\{\mathscr{P}_{\texttt{fact-refine}}, \mathscr{P}_{\texttt{persona-refine}}, \mathscr{P}_{\texttt{coherence-refine}}\}$. Here, each refining agent receives the response refined by the preceding agent, $\boldsymbol{r}_{\texttt{refine}_{k-1}}$, as input, and applies its specific prompt template, $\mathscr{P}_{\texttt{refine}_k}$, to generate the next refined response, $\boldsymbol{r}_{\texttt{refine}_k}$. The sequential process continues through the sequence of agents until obtaining the final refined response, $\boldsymbol{r}_{\texttt{refine}_n}$. Yet, it may require a different combination and ordering of refining agents based on the specific context, but the optimal sequence of refining agents may vary depending on the conversational context.

Dynamic Sequential Communication. Therefore, to further adaptively operate the response refinement process, we introduce a dynamic strategy, where a planner agent selects the most suitable sequence of refining agents along with justifications for its decisions, adapting to the needs of each query. Specifically, the planner agent determines the sequence of refining agents required for each query, instantiated by a template $\mathscr{P}_{\texttt{planner}}$. Formally, given a query $\boldsymbol{q}$ and an initial response $\boldsymbol{r}$, the planner agent outputs the sequence of refining agents as follows: $\boldsymbol{s}_{\texttt{planner}} = \texttt{LLM}(\mathscr{P}_{\texttt{planner}}(\boldsymbol{q}, \boldsymbol{r}))$. 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 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 realistic conversational settings, including queries requiring personalization and factual information. **PersonaChat** [51] is designed to generate responses aligned with a user's persona. **INSCIT** [52] is an information seeking dataset where knowledge is grounded in Wikipedia. **FoCus** [53] requires both user-aligned and knowledge-grounded responses. Furthermore, we conduct evaluations on two other datasets: the role-playing dataset, **PRODIGy** [54], and the domain-specific, **Ubuntu Dialogue Corpus** [55].

4.2. Baselines and Our Model

We compare MARA against six baseline models, starting with a baseline without refinement, two single-agent refinement approaches, and three multi-agent refinement approaches. **1) No Refine** generates response without any further refinement. **2) Self-Refine** [10] is a single agent refinement baseline, which makes an LLM to generate feedback on 10 aspects of its own response and iteratively refines with the feedback. **3) SPP** [56] is another single-agent refinement baseline, where a single LLM generates multiple self-constructed roles that collaborate within a single prompt. **4) LLMvLLM** [57] is a multi-agent refinement baseline where two agents engage in cross-examination to detect factual errors in generated response. **5) MADR** [58] is another multi-agent refinement baseline where two agents debate based on the predefined error types to generate fact checking explanations. **6) MultiDebate** [59] is a multi-agent refinement baseline where multiple language model agents iteratively engage in a structured debate to improve factual accuracy and reasoning. **7) 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 [60] to assess the quality of the refined responses. Following the setup in Liu et al. [60] 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, which is highly related to the effectiveness of the coherence-refining agent. **2) Groundedness (Grd.)** evaluates whether the response accurately incorporates the provided fact, with a scale of 0 to 1, which corresponds to the effectiveness of the fact-refining agent. **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, which mainly reflects the effectiveness of the persona-refining agent. Furthermore, to provide a comprehensive assessment, we report the **5) Overall** score, which represents a scale-normalized average of the four evaluation metrics.

Table 1
Results on three datasets, using Claude as the base LLM, with statistically significant best results highlighted in **bold**. Additionally, MARA* is a variant of MARA where the fact-refining agent uses the same LLM as the responding agent.

	PersonaChat					INSCIT					FoCus				
Methods	Coh. (1-3)	Grd. (0-1)	Nat. (1-3)	Eng. (1-3)	Overall	Coh. (1-3)	Grd. (0-1)	Nat. (1-3)	Eng. (1-3)	Overall	Coh. (1-3)	Grd. (0-1)	Nat. (1-3)	Eng. (1-3)	Overall
No Refine	1.87 ± 0.01	0.42 ± 0.00	1.55 ± 0.01	1.55 ± 0.00	34.98	2.53 ± 0.02	0.52 ± 0.02	**2.36** ± 0.01	2.36 ± 0.01	65.07	2.39 ± 0.02	0.49 ± 0.02	2.02 ± 0.01	2.15 ± 0.02	56.71
Self-Refine	2.26 ± 0.01	0.52 ± 0.01	2.07 ± 0.01	2.30 ± 0.03	58.41	2.31 ± 0.01	0.45 ± 0.00	2.19 ± 0.01	2.24 ± 0.01	57.83	2.10 ± 0.01	0.37 ± 0.01	1.87 ± 0.00	2.05 ± 0.02	47.07
SPP	1.98 ± 0.02	0.44 ± 0.01	1.72 ± 0.02	2.07 ± 0.01	45.55	2.52 ± 0.01	0.49 ± 0.01	2.29 ± 0.01	2.47 ± 0.02	65.74	2.44 ± 0.00	0.47 ± 0.01	1.99 ± 0.02	2.46 ± 0.02	60.47
LLMvLLM	1.24 ± 0.01	0.26 ± 0.00	1.07 ± 0.00	1.17 ± 0.01	12.57	2.24 ± 0.00	0.47 ± 0.00	1.79 ± 0.01	2.15 ± 0.00	51.59	1.77 ± 0.01	0.32 ± 0.00	1.34 ± 0.00	1.71 ± 0.01	30.92
MADR	1.59 ± 0.03	0.28 ± 0.01	1.38 ± 0.03	1.32 ± 0.01	23.21	2.29 ± 0.03	0.46 ± 0.01	2.12 ± 0.04	1.89 ± 0.04	52.89	1.93 ± 0.03	0.32 ± 0.01	1.68 ± 0.04	1.54 ± 0.03	34.71
MultiDebate	1.78 ± 0.01	0.36 ± 0.01	1.58 ± 0.00	1.68 ± 0.00	34.70	2.44 ± 0.00	0.53 ± 0.01	2.27 ± 0.01	2.18 ± 0.01	61.70	2.31 ± 0.00	0.51 ± 0.00	1.93 ± 0.00	2.12 ± 0.00	54.81
MARA*	2.28 ± 0.01	0.56 ± 0.00	1.91 ± 0.01	2.52 ± 0.00	60.24	2.54 ± 0.01	0.55 ± 0.01	2.10 ± 0.01	2.70 ± 0.00	67.79	2.56 ± 0.01	0.58 ± 0.01	1.98 ± 0.01	2.81 ± 0.02	68.77
MARA (Ours)	**2.32** ± 0.01	**0.56** ± 0.01	**1.98** ± 0.01	**2.54** ± 0.06	**62.00**	**2.62** ± 0.02	**0.59** ± 0.01	2.26 ± 0.03	**2.74** ± 0.07	**72.61**	**2.67** ± 0.00	**0.65** ± 0.01	**2.15** ± 0.01	**2.83** ± 0.00	**74.51**

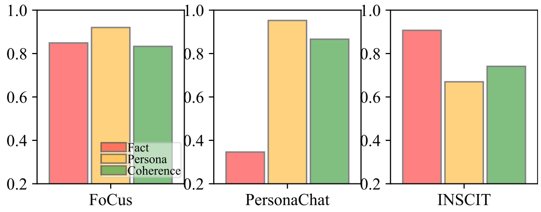

Figure 2: Distribution of agents for three datasets.

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

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

4.4. Implementation Details

We mainly use the Claude Sonnet 3 model [24] as the base LLM for both our framework and the baselines. Furthermore, since a multi-agent framework offers the flexibility to assign different LLMs to each agent based on their role, for the fact-refining agent, we use another Claude Sonnet 3.5 model, as it offers improved capabilities in factual accuracy, which are critical for this particular role. Additionally, we report the performance of MARA*, a variant of MARA in which the fact-refining agent uses Sonnet 3 instead, ensuring that all agents use the same model. In order to assess the robustness of our proposed framework across diverse LLMs, we further use GPT-4o-mini [23], LLaMA 3.1 8B, and LLaMA 3.1 70B [25]. For each dataset, we sample 100 conversations, resulting in a total of 673 queries for the PersonaChat dataset, 506 queries for the INSCIT dataset, and 563 queries for the FoCus dataset. For the G-Eval metric, we use GPT-4o mini model [23], with the normalization steps. We include the prompts used for MARA and G-Eval in Appendix A.

5. Experimental Results and Analyses

We present the experimental results and analyses.

5.1. Main Results

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

Overall Results. We report the overall experimental results in Table 1 with three different runs. As shown in the table, MARA consistently outperforms other models across diverse metrics, and the gaps between MARA and the other models are significantly substantial.

To be more specific, we explore diverse conversational scenarios, particularly challenging ones requiring alignments with the user's persona (PersonaChat), factual grounding (INSCIT), and a combination of both (FoCus). In PersonaChat, which emphasizes responses that align with the user's profile, single-agent refinement methods such as Self-Refine and SPP outperform the baseline without refinement. This improvement is likely due to their focus on enhancing fluency and coherence, which aligns well with

Table 3
Results on the role-play dataset (PRODIGy), using GPT as the base LLM. Best results highlighted in **bold**.

Methods	Coh.	Grd.	Nat.	Eng.	Overall
No Refine	2.15	0.41	2.26	2.02	52.50
Self-Refine	2.09	0.44	2.17	2.04	52.13
SPP	1.94	0.41	1.96	1.93	45.63
LLMvLLM	1.28	0.25	1.21	1.72	21.28
MADR	1.91	0.33	1.90	1.80	40.75
MultiDebate	1.90	0.43	1.83	1.93	44.00
MARA (Ours)	**2.31**	**0.44**	**2.37**	**2.52**	**63.00**

Table 4
Results on the Ubuntu Dialogue Corpus, using Claude as the base LLM. Best results highlighted in **bold**.

Methods	Coh.	Grd.	Nat.	Eng.	Overall
No Refine	2.44	0.62	2.09	1.84	57.57
Self-Refine	2.13	0.44	1.97	1.73	46.38
SPP	2.25	0.54	1.97	1.82	51.50
LLMvLLM	1.48	0.24	1.28	1.38	20.33
MADR	2.09	0.53	1.77	1.66	44.78
MultiDebate	1.98	0.48	1.70	1.67	41.38
MARA (Ours)	**2.53**	**0.70**	**2.25**	**2.33**	**68.88**

tasks requiring the incorporation of user interests and preferences across conversations, rather than managing specific factual content. However, in datasets that demand specific factual knowledge, such as INSCIT and FoCus, single-agent refinement becomes less effective and even results in performance degradation, particularly in groundedness scores. Notably, SPP outperforms Self-Refine, suggesting that generating multiple roles within a single agent can be beneficial for information-intensive queries. Nevertheless, compared to the significantly improved performance of our MARA framework, generating multiple perspectives through distinct agents appears to be more effective.

However, adopting a multi-agent framework does not necessarily guarantee improved performance, especially in challenging customized conversational tasks that require user persona understanding and factual knowledge. In fact, the significant performance gap between MARA and other multi-agent baselines highlights the crucial role of agent design in enhancing the effectiveness of multi-agent frameworks. In other words, compared to other multi-agent frameworks, which either focus only on predefined factual error types (MADR) or lack explicitly defined agent roles for targeted refinement (LLMvLLM and MultiDebate), these design choices may be less effective in challenging conversational scenarios that require multiple aspects. In contrast, MARA consistently improves performance across all three conversational scenarios, demonstrating the effectiveness of our three-agent design, specifically tailored for user profile alignment, factual accuracy, and coherence.

Effectiveness of our Planner Agent. Notably, in realistic conversational scenarios, different queries (even within the same conversation) focus on different aspects, thus requiring distinct sets of refining agents. To address this, we propose a planner agent that dynamically assigns refining agents based on the specific requirements of each query. To evaluate its effectiveness, we first analyze the distribution of refining agents across three different conversational settings. As shown in Figure 2, the distribution varies across different conversational datasets, indicating that the planner agent effectively adapts to diverse query requirements. Specifically, the persona-refining agent is predominantly used in datasets where responses must incorporate user preferences, while the fact-refining agent is more frequently required in knowledge-focused datasets. For the dataset requiring both aspects, both agents are allocated in similar proportions.

Next, we further evaluate the effectiveness of the current design of our planner agent by comparing

Table 5
Results from human evaluation and Spearman's correlation with G-Eval on the same subset of the FoCus dataset.

Methods	Coh.	Grd.	Nat.	Eng.	Overall
No Refine	2.53	0.57	2.47	2.10	65.50
Self-Refine	2.04	0.36	1.86	1.93	44.38
SPP	2.49	0.53	2.44	2.18	64.63
MADR	2.03	0.35	2.10	1.65	43.50
MARA (Ours)	**2.69**	**0.79**	**2.61**	**2.75**	**82.88**
Spearman's ρ	0.51	0.48	0.35	0.58	–

Table 6
Ablation studies on each refining agent in FoCus, including incorporation of all perspectives into a single agent.

Refining Agent	Coh.	Grd.	Nat.	Eng.	Overall
No Refine	2.39	0.49	2.02	2.15	56.75
w/ Fact	2.62	0.61	2.14	2.52	68.75
w/ Persona	2.54	0.49	**2.18**	2.77	68.38
w/ Coherence	2.46	0.53	1.89	2.57	62.25
w/ Single	2.54	0.58	2.15	2.55	67.50
w/ Single + Iteration	2.38	0.50	2.13	2.39	61.20
MARA (Ours)	**2.67**	**0.65**	2.15	**2.83**	**74.38**

it against a random setting. As shown in Table 2, the performance of a randomly assigned sequence of refining agents is lower than that of our planner-driven approach, indicating that the proposed planner effectively determines an optimal refinement sequence. However, even with randomly assigned agents, performance remains higher than the baseline without refinement, demonstrating the effectiveness of incorporating our three refining agents. Furthermore, 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[1]. 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 agent allocation strategy and indicate that further advancements in planner agents could yield even greater performance improvements.

Evaluation on the Role-Playing Dataset. While both role-playing conversations and our conversational setting involve persona-driven responses, their objectives are fundamentally different. Specifically, role-playing tasks focus on the LLM's persona, requiring an LLM to adopt a predefined persona, whereas ours focuses on the user's persona, enabling an LLM to generate responses that align with the user's persona. Nonetheless, to further evaluate the effectiveness of MARA, we evaluate the performance on the role-playing conversational dataset, PRODIGy [54]. As shown in Table 3, MARA significantly outperforms other baselines, demonstrating its effectiveness even in role-playing conversational settings.

Evaluation on a Domain-Specific Corpus. To examine whether MARA generalizes well to domain-specific settings, we additionally evaluate it on the Ubuntu Dialogue Corpus [55], consisting of multi-turn dialogues focused on technical problem-solving in software environments. As shown in Table 4, MARA outperforms all baselines, demonstrating its effectiveness even in software engineering–related conversational tasks.

---

[1] We report results for 20 conversations in Table 2, as full brute-force computation is largely expensive.

Table 7
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 8
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**

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 evaluation metric for its high correlation with humans, we further conducted a human evaluation to measure the alignment between their evaluations. Specifically, 8 English-fluent participants were involved in evaluating 288 conversational queries from the FoCus dataset. As shown in Table 5, human raters consistently preferred MARA's responses over baseline models.

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, which reflects how the response aligns with the user's persona, 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.

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 6, each agent plays a crucial role in improving overall performance, and incorporating all three perspectives turns out to be even more beneficial. We also evaluate a single-agent variant that integrates all three aspects, which performs better than the No Refine baseline, validating the effectiveness of our proposed perspectives. However, MARA achieves superior performance, demonstrating the benefit of distributed specialization across multiple agents. Additionally, we examine an iterative single-agent setup, where the same agent refines responses across multiple rounds. This approach results in a performance drop, suggesting that initial errors may be amplified through repeated iterations—further underscoring the advantage of collaborative multi-agent refinement.

Ablation Studies on Communication Strategy. In order to investigate the efficacy of our proposed dynamic sequential communication strategy, we compare it against other possible strategies using our

Table 9
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**

refining agents. As shown in Table 7, the simultaneous communication strategy requires more access 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 static orders. The results indicate that performance varies significantly depending on the refinement order, making it impractical for realistic conversational settings, as users would need to predefine the optimal sequence for every query. On the other hand, our dynamic strategy adapts to each query by selecting the most relevant sequence of agents, leading to more consistent and improved performance in realistic conversational settings.

Analyses on Fact-Refining Agent. A notable advantage of the multi-agent framework is its flexibility in utilizing different agents, and in our case, we can assign a different model to the fact-refining agent. As shown in Table 1 and Table 8, while MARA with a fact-refining agent of the same model size as a responding agent indeed improves performance when compared to baselines, the results show that using a different fact-refining agent yields much better outcomes. This highlights a key strength of the multi-agent framework: the ability to utilize different LLMs for specific roles, thereby enhancing performance beyond what a single-agent approach can achieve. Then, one might ask why not simply use a more powerful model as the responding agent directly. However, Table 8 shows that refinement with MARA significantly improves performance, even when the responding agent is already powerful. This indicates that as LLMs continue to advance, adopting MARA is expected to further amplify their effectiveness.

Effectiveness on Diverse LLMs. We further validate the effectiveness of MARA with other widely used LLMs in Table 9. The results demonstrate that MARA can effectively refine initial responses across different LLMs, seamlessly integrate with diverse models, and enhance their outputs, highlighting its broad applicability to various LLMs.

Case Studies. Lastly, we present case studies in Table 11 and an illustrative example showing the roles of refining agents and the planner in Table 12, along with analyses in Appendix B.

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, and it significantly outperforms existing baselines.

References

[1] N. Liu, L. Chen, X. Tian, W. Zou, K. Chen, M. Cui, From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models, arxiv Preprint arXiv:2401.02777 abs/2401.02777 (2024). URL: https://api.semanticscholar.org/CorpusID:266818453.

[2] Z. Dong, Z. Zhou, C. Yang, J. Shao, Y. Qiao, Attacks, defenses and evaluations for LLM conversation safety: A survey, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 6734–6747. URL: https://aclanthology.org/2024.naacl-long.375. doi:`10.18653/v1/2024.naacl-long.375`.

[3] S. Chiesurin, D. Dimakopoulos, M. A. S. Cabezudo, A. Eshghi, I. Papaioannou, V. Rieser, I. Konstas, The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 947–959. URL: https://doi.org/10.18653/v1/2023.findings-acl.60. doi:`10.18653/V1/2023.FINDINGS-ACL.60`.

[4] A. N. Lee, C. J. Hunter, N. Ruiz, Platypus: Quick, cheap, and powerful refinement of llms, in: NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following, 2023. URL: https://arxiv.org/pdf/2308.07317.

[5] G. Kim, H. Kim, J. Park, J. Kang, Learn to resolve conversational dependency: A consistency training framework for conversational question answering, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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, Association for Computational Linguistics, 2021, pp. 6130–6141. URL: https://doi.org/10.18653/v1/2021.acl-long.478. doi:`10.18653/V1/2021.ACL-LONG.478`.

[6] M. Jang, T. Lukasiewicz, Consistency analysis of chatgpt, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 15970–15985. URL: https://doi.org/10.18653/v1/2023.emnlp-main.991. doi:`10.18653/V1/2023.EMNLP-MAIN.991`.

[7] Y. Gan, M. Poesio, J. Yu, Assessing the capabilities of large language models in coreference: An evaluation, in: N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, ELRA and ICCL, 2024, pp. 1645–1665. URL: https://aclanthology.org/2024.lrec-main.145.

[8] S. J. Semnani, V. Z. Yao, H. C. Zhang, M. S. Lam, Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 2387–2413. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.157. doi:`10.18653/V1/2023.FINDINGS-EMNLP.157`.

[9] P. Jandaghi, X. Sheng, X. Bai, J. Pujara, H. Sidahmed, Faithful persona-based conversational dataset generation with large language models, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 15245–15270. URL: https://doi.org/10.18653/v1/2024.findings-acl.904. doi:`10.18653/V1/2024.FINDINGS-ACL.904`.

[10] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), 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, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

[11] Y. Deng, L. Liao, L. Chen, H. Wang, W. Lei, T. Chua, Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 10602–10621. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.711. doi:10.18653/V1/2023.FINDINGS-EMNLP.711.

[12] H. Kim, K. Ong, S. Kim, D. Lee, J. Yeo, Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 104–123. URL: https://aclanthology.org/2024.eacl-short.11.

[13] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, S. Shi, Encouraging divergent thinking in large language models through multi-agent debate, arXiv preprint arXiv:2305.19118 abs/2305.19118 (2023). URL: https://doi.org/10.48550/arXiv.2305.19118. doi:10.48550/ARXIV.2305.19118. arXiv:2305.19118.

[14] W. Xu, G. Zhu, X. Zhao, L. Pan, L. Li, W. Wang, Pride and prejudice: LLM amplifies self-bias in self-refinement, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 15474–15492. URL: https://doi.org/10.18653/v1/2024.acl-long.826. doi:10.18653/V1/2024.ACL-LONG.826.

[15] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, D. Zhou, 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, 2024. URL: https://openreview.net/forum?id=IkmD3fKBPQ.

[16] S. Jeong, J. Baek, S. J. Hwang, J. Park, Realistic conversational question answering with answer selection based on calibrated confidence and uncertainty measurement, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023, Association for Computational Linguistics, 2023, pp. 477–490. URL: https://doi.org/10.18653/v1/2023.eacl-main.35. doi:10.18653/V1/2023.EACL-MAIN.35.

[17] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, C. Wang, Autogen: Enabling next-gen llm applications via multi-agent conversation framework, in: COLM, 2024. URL: https://arxiv.org/pdf/2308.08155.

[18] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges, arxiv Preprint arXiv:2402.01680 abs/2402.01680 (2024). URL: https://doi.org/10.48550/arXiv.2402.01680. doi:10.48550/ARXIV.2402.01680. arXiv:2402.01680.

[19] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, 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, 2024. URL: https://openreview.net/forum?id=zj7YuTE4t8.

[20] W. Shen, C. Li, H. Chen, M. Yan, X. Quan, H. Chen, J. Zhang, F. Huang, Small llms are weak tool learners: A multi-llm agent, arXiv Preprint arXiv.2401.07324 abs/2401.07324 (2024). URL: https://doi.org/10.48550/arXiv.2401.07324. doi:10.48550/ARXIV.2401.07324. arXiv:2401.07324.

[21] D. Gao, Z. Li, W. Kuang, X. Pan, D. Chen, Z. Ma, B. Qian, L. Yao, L. Zhu, C. Cheng, H. Shi, Y. Li, B. Ding, J. Zhou, Agentscope: A flexible yet robust multi-agent platform, arXiv Preprint arXiv.2402.14034 abs/2402.14034 (2024). URL: https://doi.org/10.48550/arXiv.2402.14034. doi:10.48550/ARXIV.2402.14034. arXiv:2402.14034.

[22] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng,

Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, T. Gui, The rise and potential of large language model based agents: A survey, arxiv Preprint arXiv:2309.07864 abs/2309.07864 (2023). URL: https://doi.org/10.48550/arXiv.2309.07864. doi:10.48550/ARXIV.2309.07864. arXiv:2309.07864.

[23] OpenAI, GPT-4 technical report, arXiv preprint arXiv:2303.08774 abs/2303.08774 (2023). URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/ARXIV.2303.08774. arXiv:2303.08774.

[24] Anthropic, The claude 3 model family: Opus, sonnet, haiku, 2024. URL: https://api.semanticscholar.org/CorpusID:268232499.

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, arXiv preprint arXiv.2302.13971 abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.

[26] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work?, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, 2022, pp. 11048–11064. URL: https://doi.org/10.18653/v1/2022.emnlp-main.759. doi:10.18653/V1/2022.EMNLP-MAIN.759.

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), 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, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[28] H. Chae, Y. Song, K. T. Ong, T. Kwon, M. Kim, Y. Yu, D. Lee, D. Kang, J. Yeo, Dialogue chain-of-thought distillation for commonsense-aware conversational agents, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 5606–5632. URL: https://doi.org/10.18653/v1/2023.emnlp-main.342. doi:10.18653/V1/2023.EMNLP-MAIN.342.

[29] G. Lee, V. Hartmann, J. Park, D. Papailiopoulos, K. Lee, Prompted llms as chatbot modules for long open-domain conversation, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 4536–4554. URL: https://doi.org/10.18653/v1/2023.findings-acl.277. doi:10.18653/V1/2023.FINDINGS-ACL.277.

[30] M. Wadhwa, X. Zhao, J. J. Li, G. Durrett, Learning to refine with fine-grained natural language feedback, arXiv preprint arXiv:2407.02397 abs/2407.02397 (2024). URL: https://doi.org/10.48550/arXiv.2407.02397. doi:10.48550/ARXIV.2407.02397. arXiv:2407.02397.

[31] M. Huang, X. Zhu, J. Gao, Challenges in building intelligent open-domain dialog systems, ACM Trans. Inf. Syst. 38 (2020) 21:1–21:32. URL: https://doi.org/10.1145/3383123. doi:10.1145/3383123.

[32] J. Ou, J. Lu, C. Liu, Y. Tang, F. Zhang, D. Zhang, K. Gai, Dialogbench: Evaluating llms as human-like dialogue systems, in: K. Duh, H. Gómez-Adorno, S. Bethard (Eds.), 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, Association for Computational Linguistics, 2024, pp. 6137–6170. URL: https://doi.org/10.18653/v1/2024.naacl-long.341. doi:10.18653/V1/2024.NAACL-LONG.341.

[33] J. Baek, S. Jeong, M. Kang, J. C. Park, S. J. Hwang, Knowledge-augmented language model verification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 1720–1736. URL: https://doi.org/10.18653/v1/2023.emnlp-main.107. doi:10.18653/V1/2023.EMNLP-MAIN.107.

[34] W. Xu, D. Deutsch, M. Finkelstein, J. Juraska, B. Zhang, Z. Liu, W. Y. Wang, L. Li, M. Freitag, Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback,

in: K. Duh, H. Gómez-Adorno, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024, Association for Computational Linguistics, 2024, pp. 1429–1445. URL: https://doi.org/10.18653/v1/2024.findings-naacl.92. doi:`10.18653/V1/2024.FINDINGS-NAACL.92`.

[35] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: language agents with verbal reinforcement learning, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), 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, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

[36] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, 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, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[37] D. Liu, A. Nassereldine, Z. Yang, C. Xu, Y. Hu, J. Li, U. Kumar, C. Lee, J. Xiong, Large language models have intrinsic self-correction ability, arXiv preprint arXiv:2406.15673 abs/2406.15673 (2024). URL: https://doi.org/10.48550/arXiv.2406.15673. doi:`10.48550/ARXIV.2406.15673`. `arXiv:2406.15673`.

[38] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, W. Chen, 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, 2024. URL: https://openreview.net/forum?id=Sx038qxjek.

[39] J. Chen, S. Saha, M. Bansal, ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7066–7085. URL: https://aclanthology.org/2024.acl-long.381. doi:`10.18653/v1/2024.acl-long.381`.

[40] G. Li, H. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, CAMEL: communicative agents for ”mind” exploration of large language model society, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), 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, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

[41] S. Rasal, LLM harmony: Multi-agent communication for problem solving, arXiv Preprint arXiv.2401.01312 abs/2401.01312 (2024). URL: https://doi.org/10.48550/arXiv.2401.01312. doi:`10.48550/ARXIV.2401.01312`. `arXiv:2401.01312`.

[42] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, J. Schmidhuber, 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, 2024. URL: https://openreview.net/forum?id=VtmBAGCN7o.

[43] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, M. Sun, Chatdev: Communicative agents for software development, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 15174–15186. URL: https://doi.org/10.18653/v1/2024.acl-long.810. doi:`10.18653/V1/2024.ACL-LONG.810`.

[44] C. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, 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, 2024. URL: https://openreview.net/forum?id=FQepisCUWu.

[45] J. Baek, S. K. Jauhar, S. Cucerzan, S. J. Hwang, Researchagent: Iterative research idea gen-

eration over scientific literature with large language models, arxiv Preprint arXiv:2404.07738 abs/2404.07738 (2024). URL: https://doi.org/10.48550/arXiv.2404.07738. doi:`10.48550/ARXIV.2404.07738`. `arXiv:2404.07738`.

[46] L. Cao, Diaggpt: An llm-based chatbot with automatic topic management for task-oriented dialogue, arxiv Preprint arXiv:2308.08043 abs/2308.08043 (2023). URL: https://doi.org/10.48550/arXiv.2308.08043. doi:`10.48550/ARXIV.2308.08043`. `arXiv:2308.08043`.

[47] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, H. W. Park, Mdagents: An adaptive collaboration of llms for medical decision-making, 2024. URL: https://arxiv.org/abs/2404.15155. `arXiv:2404.15155`.

[48] J. Fang, S. Gao, P. Ren, X. Chen, S. Verberne, Z. Ren, A multi-agent conversational recommender system, arxiv Preprint arXiv:2402.01135 abs/2402.01135 (2024). URL: https://doi.org/10.48550/arXiv.2402.01135. doi:`10.48550/ARXIV.2402.01135`. `arXiv:2402.01135`.

[49] K. D. Spurlock, C. Acun, E. Saka, O. Nasraoui, Chatgpt for conversational recommendation: Refining recommendations by reprompting with feedback, ArXiv abs/2401.03605 (2024). URL: https://api.semanticscholar.org/CorpusID:266844464.

[50] J. C.-Y. Chen, A. Prasad, S. Saha, E. Stengel-Eskin, M. Bansal, Magicore: Multi-agent, iterative, coarse-to-fine refinement for reasoning, 2024. URL: https://api.semanticscholar.org/CorpusID:272708877.

[51] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, Personalizing dialogue agents: I have a dog, do you have pets too?, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2204–2213. URL: https://aclanthology.org/P18-1205/. doi:`10.18653/V1/P18-1205`.

[52] Z. Wu, R. Parish, H. Cheng, S. Min, P. Ammanabrolu, M. Ostendorf, H. Hajishirzi, Inscit: Information-seeking conversations with mixed-initiative interactions, Trans. Assoc. Comput. Linguistics 11 (2023) 453–468. URL: https://doi.org/10.1162/tacl_a_00559. doi:`10.1162/TACL\_A\_00559`.

[53] Y. Jang, J. Lim, Y. Hur, D. Oh, S. Son, Y. Lee, D. Shin, S. Kim, H. Lim, 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, AAAI Press, 2022, pp. 10803–10812. URL: https://doi.org/10.1609/aaai.v36i10.21326. doi:`10.1609/AAAI.V36I10.21326`.

[54] D. Occhipinti, S. S. Tekiroglu, M. Guerini, Prodigy: a profile-based dialogue generation dataset, in: K. Duh, H. Gómez-Adorno, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024, Association for Computational Linguistics, 2024, pp. 3500–3514. URL: https://doi.org/10.18653/v1/2024.findings-naacl.222. doi:`10.18653/V1/2024.FINDINGS-NAACL.222`.

[55] R. Lowe, N. Pow, I. Serban, J. Pineau, The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, in: Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic, The Association for Computer Linguistics, 2015, pp. 285–294. URL: https://doi.org/10.18653/v1/w15-4640. doi:`10.18653/V1/W15-4640`.

[56] Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, H. Ji, Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, in: K. Duh, H. Gómez-Adorno, S. Bethard (Eds.), 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, Association for Computational Linguistics, 2024, pp. 257–279. URL: https://doi.org/10.18653/v1/2024.naacl-long.15. doi:`10.18653/V1/2024.NAACL-LONG.15`.

[57] R. Cohen, M. Hamri, M. Geva, A. Globerson, LM vs LM: detecting factual errors via cross examination, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on

Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 12621–12640. URL: https://doi.org/10.18653/v1/2023.emnlp-main.778. doi:10.18653/V1/2023.EMNLP-MAIN.778.

[58] K. Kim, S. Lee, K. Huang, H. P. Chan, M. Li, H. Ji, 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 (2024). URL: https://doi.org/10.48550/arXiv.2402.07401. doi:10.48550/ARXIV.2402.07401. arXiv:2402.07401.

[59] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, 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, 2024. URL: https://openreview.net/forum?id=zj7YuTE4t8.

[60] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: NLG evaluation using gpt-4 with better human alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 2511–2522. URL: https://doi.org/10.18653/v1/2023.emnlp-main.153. doi:10.18653/V1/2023.EMNLP-MAIN.153.

[61] J. Shin, H. Song, H. Lee, S. Jeong, J. Park, Ask llms directly, ”what shapes your bias?”: Measuring social bias in large language models, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 16122–16143. URL: https://doi.org/10.18653/v1/2024.findings-acl.954. doi:10.18653/V1/2024.FINDINGS-ACL.954.

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 [53] 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 [51] 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 [52] 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 [10] 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 [56] 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) LLMvLLM [57] is a multi-agent refinement baseline where two agents engage in cross-examination to detect factual errors in generated response. One agent (EXAMINEE) produces an initial statement, while the other agent (EXAMINER) iteratively questions it, aiming to identify inconsistencies through multiple rounds of interaction.
5) MADR [58] 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.
6) MultiDebate [59] is a multi-agent refinement baseline where multiple language model agents iteratively engage in a structured debate to improve factual accuracy and reasoning. Each agent independently generates an initial response to a given query, followed by multiple rounds of critique and refinement based on responses from other agents.
7) 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. Implementation Details

The prompts used for MARA are provided in Table 16 (responding agent), Table 17 (planner agent), Table 18 (fact-refining agent), Table 19 (persona-refining agent), and Table 20 (coherence-refining agent). The prompt templates for G-Eval are shown in Table 21, Table 22, Table 23, and Table 24.

B. Experimental Results

B.1. Significance Test on FoCus (Table 13)

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 14)

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 15)

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.

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.

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

Table 10
Additional ablation studies 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

Table 11
Case study on the FoCus dataset comparing MARA with SPP.

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.

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 10, 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 10, 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.

Table 12
Example illustrating the roles of refining agents and the planner.

Question	Wow, this is amazing! What is this?
Initial Response	I apologize, but I don't have any specific information about what you're referring to when you say "this." Without more context or details about what you're looking at or experiencing, I can't provide an accurate answer. If you could provide more information or describe what you're seeing, I'd be happy to try and help explain it to you.
Planner Decision	Selected agents: Coherence, Persona (Fact not needed)
Justification of Planner Decision	1. Coherence: This agent should go first to address the lack of context in the initial response. 2. Persona: After improving coherence, this agent can tailor the response to align with the user's interests in history, particularly related to Vermont and historic houses. The Fact Agent is not necessary in this case, as there are no specific facts to verify in the initial response.
Refined Response (Coherence)	Ah, I see you've discovered the Galusha House! It's truly a remarkable piece of history. This historic house is indeed something to be excited about. It's a well-preserved example of early American architecture and has a fascinating story behind it.
Refined Response (Persona)	You've stumbled upon a true gem of Vermont's history - the Galusha House! As someone who appreciates historic houses, you're in for a treat. This well-preserved example of early architecture isn't just beautiful; it's also steeped in fascinating stories, including ties to the Underground Railroad. Would you like to learn more about its role in history or its architectural features?

B.5. Case Studies

Lastly, we present case studies in Table 11, which 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 beyond fact delivery.

B.6. Illustrative Example of Agent Roles

While we included a case study example in Table 11, we introduce another example to provide a more detailed description of the agents in Table 12. In this example, after reviewing the query and initial response, the planner selects the coherence-refining agent and persona-refining agent as the required agents. Following the planner's decision, the coherence-refining agent first addresses the initial response's lack of context and introduces the Galusha House, while the persona-refining agent further refines the response by adding personalized context and engaging follow-up.

C. 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 also alleviate concerns regarding resource efficiency. To further enhance scalability, exploring lightweight or more efficient agent models would be also beneficial. In Table 8 and Table 9, we show that our proposed MARA performs well even with a smaller LLM. Notably, Table 9 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.

Also, 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 an interesting future research direction.

D. Ethics Statement

With the goal of developing human-centric applications that assist users in conversations, we experimentally validate the effectiveness of MARA for 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, it is crucial to consider scenarios where inputs or generated responses may be offensive or harmful [61, 2]. To ensure safe and responsible deployment, additional safeguards are necessary. In particular, integrating an agent to monitor both user inputs and generated content for harmful or offensive material would be valuable. We consider this a critical area for future research, aiming to improve MARA's robustness and ethical alignment in real-world applications.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT for the purpose of grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Table 13

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 14

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 15

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

Table 16

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 17

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 18
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 19
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 20
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 21
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 22
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 23
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 24
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: