PANDA: Persona Attributes Navigation for Detecting and Alleviating Overuse Problem in Large Language Models

Jinsung Kim*, **Seonmin Koo***, **Heuiseok Lim**†
Department of Computer Science and Engineering, Korea University
{jin62304,fhdahd,limhseok}@korea.ac.kr

Abstract

In the persona-grounded dialogue (PGD) task, it is required not only to respond fluently, but also to ground the attributes according to the current conversation topic properly. However, due to their tendency to overly ground given attributes, LLMs often generate unnatural responses provoked by using attributes that deviate from the flow of the conversation or by exploiting too many attributes at once. We term this phenomenon the *overuse* problem of LLMs. Unfortunately, research devising precise criteria and frameworks to quantitatively verify LLMs' *overuse* problem is obviously insufficient. To address this issue, we propose **P**ersona **A**ttributes **N**avigation for **D**etecting and **A**lleviating the *overuse* problem (**PANDA**) framework. **PANDA** is the first study to quantify the persona *overuse* problem of LLMs by establishing clear standards of the problem and verifying various LLMs based on them. Moreover, this framework navigates us into understanding persona attributes by introducing diverse and detailed dialogue topics that consider practical conversation situations. We provide insights related to LLMs' persona attribute *overuse* problem through comprehensive verification and analysis with **PANDA** in the PGD task. Our code and resources can be found at http://github.com/jin62304/PANDA.

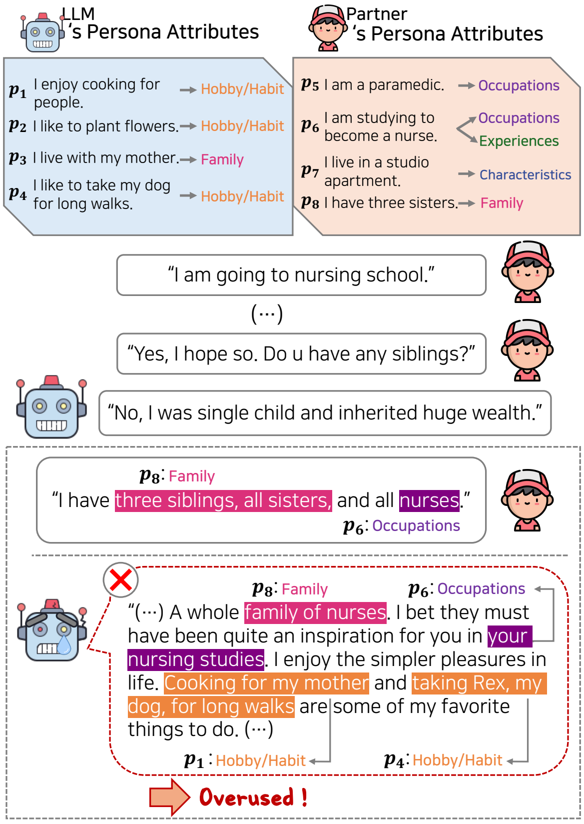

Figure 1: Example response by Mistral in the persona-based dialogue situation. LLMs often overly utilize the given persona attributes that deviate from the dialogue topics. Each color indicates a corresponding topic.

1 Introduction

The persona-grounded dialogue (PGD) task aims to generate responses that fit the conversation's flow based on pre-defined personas. This task is practical because it is effectively used in the real world, for example, empathetic chatbots in mental health care or education (De Gennaro et al., 2020; Pu et al., 2024). For the PGD task, the model is required to be able to continue a conversation appropriate to the context while appropriately utilizing its own persona (self persona) and the persona of the conversation partner (partner persona).

Previously, relatively moderate-sized pre-trained language models (PLMs) have suffered from a lack of consistent personality regarding persona grounding, such as not actively leveraging persona attributes (Li et al., 2016; Zhao et al., 2023; Li et al., 2023). This challenge has led to the revitalization of several existing studies focusing on how faithfully persona attributes are grounded. For example, there

* Equally contributed.
† Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12005–12026
November 12-16, 2024

have been attempts to consider how diligently a model grounds persona attributes, based on the cosine similarity or F1 score between given persona attributes and the model's response (Song et al., 2019; Nargund et al., 2022; Kwon et al., 2023).

Conversely, with the advent of the era of large language models (LLMs), which have greatly improved the overall fluency of responses and instruction-following abilities, a phenomenon is often observed in which LLMs tend to excessively utilize persona attributes in conversational situations, generating unnatural responses. We term this phenomenon the *overuse* problem, where the model grounds overfull persona attributes at once or exploits those that deviate from the dialogue topic. As shown in Figure 1, which contains the actual LLM's response, the LLM recklessly uses several personas, including off-topic attributes from the conversation context in the partner's utterance. This *overuse* phenomenon hinders the interaction between the system and the user in actual conversation situations and reduces end-user satisfaction, so it needs to be detected and alleviated.

Considering examples from a similar research field, numerous studies on emotional support conversation (ESC), particularly those involving role-playing tasks such as the PGD task—a significant area of research in real-world mental health care—have regarded problems that deviate from behavioral strategies appropriate for specific stages of the conversation as critical. This is due to the structured roles that participants are expected to fulfill at each stage (Kang et al., 2024; Deng et al., 2023b). However, there is insufficient research to quantify and formalize the problem and verify the status of LLMs to alleviate the persona attributes *overuse* phenomenon of LLMs in the PGD task. This is because existing persona attributes grounding studies focus on being as similar to persona attributes as possible and grounding them to the fullest extent, which is the opposite direction of the *overuse* phenomenon.

To address this issue, this study is conducted with the following research question:

- RQ) *Do LLMs overuse persona attributes in conversation situations? How can this be verified?*

Accordingly, we propose **P**ersona **A**ttributes **N**avigation for **D**etecting and **A**lleviating *overuse* problem (**PANDA**), a verification framework with a dialogue flow-aware task and measurement method that becomes the basis for alleviation by detecting LLMs' persona *overuse* problem. Moreover, we devise a verification resource construction method to execute the framework. The **PANDA** framework consists of multiple steps: dialogue labeling, persona-topic mapping, and overuse measurement.

In this study, the following phases for **PANDA** framework design are explained: i) Two criteria for quantifying the *overuse* problem–'Off-topic' and 'Excess of quantity'– are clearly set (§ 2.2), and ii) to typify persona attributes, from the dialogue topic perspective, 14 fine-grained topics are introduced that take into account the actual conversation situation. (§ 3.1). Afterward, based on i) and ii), iii) we design a task based on multiple steps for detection of the *overuse* problem (§ 3.3). iv) Additionally, the creation method of the verification dataset to perform this task is explained (§ 4.1).

Our contributions are threefold: (1) We conduct the first comprehensive study to formalize the *overuse* problem that occurs in persona-grounded dialogue tasks and verify various LLMs based on this. (2) We propose a **PANDA** framework that detects and verifies the *overuse* problem and a resource creation method for it. (3) We serve as a navigator to improve the interpretability of LLMs' use of persona attributes based on **PANDA**'s fine-grained topic taxonomy and sophisticated quantification that considers real-world conversation situations.

2 Why PANDA?

In this section, we describe why **PANDA** is significant for verifying the *overuse* phenomenon and set the criteria to quantify the *overuse* problem.

2.1 Importance of PANDA

Firstly, in real-world conversations, LLMs need the capability to use personas appropriately, taking into account the flow of the conversation, beyond simply reflecting the persona in the response. However, due to significant improvements in overall generation fluency and instruction-following ability, LLMs often generate unnatural responses that exploit a given persona unconditionally, resulting in decreased end-user satisfaction. Therefore, a tool such as **PANDA** is required to verify whether persona attributes are appropriately used considering the conversation topic of the other speaker.

Second, existing evaluation methods for the PGD tasks are insufficient to address the *overuse* problem of LLMs. Existing studies have mainly

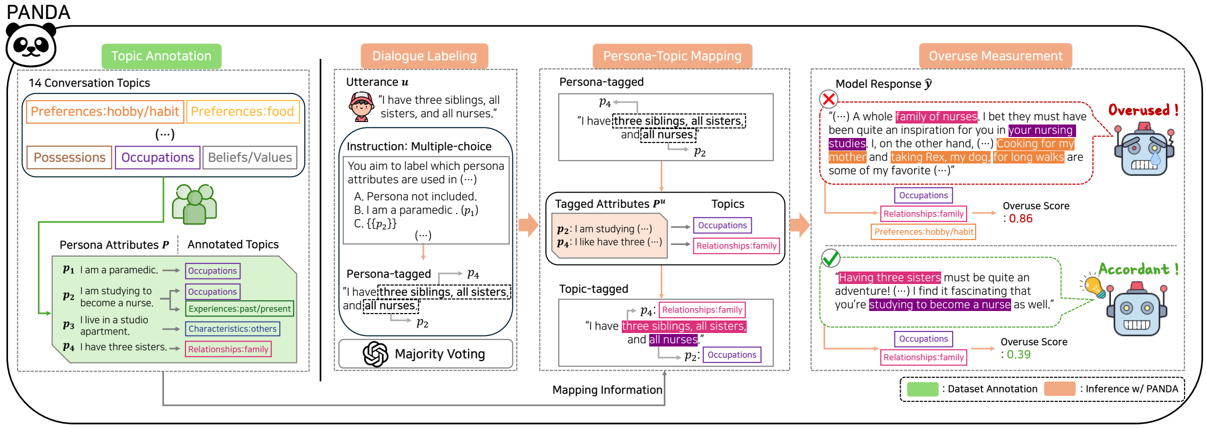

Figure 2: Overview of the designed persona overuse verification framework (**PANDA**). The overuse detection cases demonstrate a comparison of the actual generation results of Mistral (upper part) and ChatGPT (lower part) for the identical utterance of the other interlocutor.

focused on measuring model fluency by adopting scoring methods such as Rouge-L (Lin, 2004) and chrF++ (Popović, 2017). In addition, the method of evaluating the model's grounding simply based on the degree of similarity or overlapping between attributes and the model response has limitations concerning the persona *overuse* problem. That is, regarding the *overuse* problem, rather than assessing whether numerous personas are grounded regardless of the conversation topic, a practical measuring approach such as **PANDA** is needed, which complexly considers the personas included in the other person's utterances and LLMs' responses.

Thirdly, when categorizing persona attributes, defining types in detail is important for specific diagnoses and improving the model's explainability. The more diverse and clear the standards presented, the more helpful it is to improve LLM. Therefore, **PANDA** increases explainability for improving LLM capabilities through typification of persona attributes that reflect conversation situations in the real world from the perspective of the dialogue topic. By presenting approximately 2.8 times more fine-grained topics than existing studies that categorize attributes (Gao et al., 2023), the interpretability of *overuse* detection considering the conversation context is improved.

2.2 Setting Criteria for Defining Overuse

To quantify and define the *overuse* problem, this work sets two criteria for what constitutes overuse: 1) 'Off-topic' and 2) 'Excess of quantity.' To set the criteria, we took inspiration from Gricean maxims and adopted two of the four maxims, The maxims of Quantity and Relation, considering the aim and scope of this study (Grice, 1975). First, 'Off-topic' indicates using attributes that do not fit the topic of the other person's utterance. An off-topic response generated by grounding not only the topic the other speaker is dealing with, but also unrelated topics is an element that must be considered for human-like conversation (Ghazvininejad et al., 2018; Harrison et al., 2020).

Also, 'Excess of quantity' refers to generating a response on the same topic as the other speaker, but using immoderate numbers of attributes, making the conversation unnatural. A large number of retrieved results in the LLMs' responses may introduce additional noises (Wang et al., 2023). Although this has been relatively less discussed in the dialogue task, it is necessary for improving end-user satisfaction through natural conversation. We emphasize that off-topic and excess of quantity are not static and are measured taking into account the fact that the *overuse* of the same LLMs' response varies depending on the conversation context.

3 🐼 PANDA: How to Measure Persona Attribute Overuse Problem

Figure 2 demonstrates the entire overview of **PANDA**, the verification framework we propose. **PANDA** consists of dialogue labeling, persona-topic mapping, and overuse measurement tasks (orange boxes on the right). Additionally, this study includes the phase of annotating resources to acquire mapping features used for retrieval in the persona-topic mapping task (green box on the left).

Topics	Definitions	Examples
Preferences:hobby/habit	Topics about an individual's preferences regarding hobbies, habits, and voluntarily maintained periodic routines.	"I garden every weekend."
Preferences:food	The topic of personal preferences in food.	"I am a vegetarian."
Preferences:others	Topics on preferences for things other than habit and food, such as technology, animals, and plants.	"I like giant pandas."
Characteristics:appearance	Topics related to the physical characteristics of an individual.	"I have brown eyes."
Characteristics:personality	Topics related to an individual's personality traits.	"I am timid.", "I am pretty smart."
Characteristics:others	Individual's other personal features such as nationality, age, voice, and phobia.	"I live in Shanghai.", "I am good at cooking."
Relationships:family	Topics related to an individual's family relationships	"I have two older sisters."
Relationships:friend	Topics related to an individual's friends and friendships.	"I do business with my best friend."
Relationships:others	Topics related to an individual's social relationships outside of family and friends.	"I am in a team with five coworkers."
Experiences:past/present	Topics related to an individual's past experiences or current experiences.	"I was a class president during my school days."
Experiences:future/plan	Topics about an individual's plans and goals that are feasible and concrete.	"I am going to California this vacation."
Possessions	Topics related to tangible assets possessed externally by an individual rather than the individual's external/internal characteristics.	"I have a red Ferrari."
Occupations	Topics on an individual's professional status, position, and professional conduct.	"I am a lawyer.", "I am the team leader.", "I am good at giving injections to my patients."
Beliefs/Values	Attributes consisting of abstract wish verbs (wish, hope, believe, etc.), such as desires and religious beliefs. It is distinguished from 'experiences:future/plan' in feasibility and specificity.	"I want to meet my deceased grandmother in a dream.", "I am a Catholic."

Table 1: Taxonomy of dialogue topics introduced for the *overuse* problem detection.

3.1 Introducing Taxonomy of Topics

This section defines deliberately fine-grained dialogue topic taxonomy for persona attributes. Defining the taxonomy of the topics in detail helps provide the interpretation of the interaction between the user and model (Cheng et al., 2023; An and Rudinger, 2023). However, existing persona datasets lack definitions and descriptions of types that categorize persona attributes, or provide only a relatively small number of types, thereby achieving insufficient explainability for model behaviors (Zhang et al., 2018).

Therefore, inspired by Gao et al. (2023), which categorizes persona types based on studies of interactions and behavioral analysis from the perspective of user and model interplay (Dunbar et al., 1997; Cooper, 1999; Mulder and Yaar, 2006; Cooper et al., 2007), we define a fine-grained topic taxonomy for persona attributes, considering real-world dialogue situations. We expand the taxonomy by defining additional topics, such as possessions, beliefs/values, etc., that existing research has not addressed. Thus, we introduce more clarified and diverse classification standards, with a total of 14 detailed dialogue topics to categorize persona attribute instances.

Table 1 demonstrates the definition descriptions and corresponding examples of our devised dialogue topics. Including individual preferences in daily life and characteristics that signify personal traits, we introduce topics on persona attributes with approximately 2.8 times more definitions than existing studies. Therefore, our proposed taxonomy of topics can quantitatively assess persona overuses and enhance the interpretability of model responses.

3.2 Denotation: Persona-grounded Dialogue Generation Task

Before quantifying the *overuse* problem, we denote the PGD task. When the entire dialogue between two speakers is D, the other person's last utterance for generating the model's response is denoted u, and the history set up to just before u is denoted H. In other words, the relation between D, H, and u is as follows: $D = \{u_1, u_2, ..., u_{|D|}\}$, $H = \{u_1, u_2, ..., u_{(|D|-1)}\}$, and $u = u_{|D|}$. Additionally, when P is the set of persona attributes corresponding to the dialogue D, the individual element within it is p_i. Therefore, given a dialogue D and a persona attribute set P, the model response generated by LLM is denoted as $\hat{y}$. Based on the model's response generated in this way, it is verified whether there is an *overuse* problem.

3.3 Problem Formalization: Persona Overuse

Let T be a set of dialogue topics whose components are individual topics t defined in Section 3.1. That is, $T = \{$`[preferences:hobby/habit]`, `[preferences:food]`$, \cdots\}$. We quantify the persona attributes *overuse* problem as the cases where 1) Off-topic or 2) Excess of quantity (§ 2.2) occurs in the model's response $\hat{y}$, when compared to the dialogue topics included in the partner interlocutor's last utterance u.

Accordingly, to obtain the set of topics included in u and $\hat{y}$, the following two steps are taken: dialogue labeling and persona–topic mapping. In other words, the relation between u (or $\hat{y}$) and T is calculated through 1) operations on the relation between u (or $\hat{y}$) and the set P of persona attributes, and 2) operations on the relation between P and the set T of topics.

Dialogue Labeling. The persona attributes labeling function $TAG_P(\cdot)$ labels the set of persona attributes P^x contained in any given text x as follows:

$$\begin{aligned} TAG_P(x) &= P^x \\ &= \{p_i | p_i \in P \wedge p_i \in x\}_{p_i \in P,\, i \in \{0,1,\cdots,|P|\}} \end{aligned} \tag{1}$$

A subset P^u, $P^{\hat{y}}$ consisting of persona attributes $p_i \in P$ is composed, respectively from given u ($\in D$) and $\hat{y}$ in the PGD task.

Persona–Topic Mapping. The persona-topic mapping function $TAG_T(\cdot)$ outputs the set of topics T^{P^x} included in the given attributes set P^x through Equation. 1. First, this function performs mapping to topics for each element p_i^x included in P^x as follows:

$$\begin{aligned} \underset{p_i^x \in P^x}{TAG_T(p_i^x)} &= T_i^{P^x} \\ &= \{t_j | t_j \in T \wedge t_j \in p_i^x\}_{i \in \{0,1,\cdots,|P^x|\},\, j \in \{0,1,\ldots,|T|\}} \end{aligned} \tag{2}$$

Afterward, by aggregating the topic set $T_i^{P^x}$ that corresponds to p_i^x calculated in Equation 2, T^{P^x} is computed as follows:

$$TAG_T(P^x) = T^{P^x} = \sum_{i=1}^{|P^x|} T_i^{P^x} \tag{3}$$

Suppose $P^u = [p_1, p_2]$, and the set of topics corresponding to p_1 is $T_1^{P^u} = [t_1, t_2]$ and p_2 is $T_2^{P^u} = [t_1, t_3]$, respectively. Then, $T^{P^u} = [t_1, t_1, t_2, t_3]$.

To summarize, computing the relation between given textual data x and a set of topics T involves the following two steps: 1) P^x is calculated through the persona tagging (dialogue labeling) step of text x. 2) The element p_i^x of P^x is mapped to the set $T_i^{p_i^x}$ consisting of elements of T, and the sets are merged to construct T^{P^x}. That is, $T^{P^x} = TAG_T(TAG_P(x))$. Therefore, the final sets of dialogue topics for the response $\hat{y}$ generated by the model and the counterpart's last utterance u are $T^{P^{\hat{y}}}$ and T^{P^u}. Also, the number of these sets is $|T^{P^{\hat{y}}}|$, $|T^{P^u}|$, respectively.

Overuse Measurement. As described in Section 3.2, the entire dialogue D (including the dialogue history H and the other interlocutor's last utterance u) and the persona set P are assigned to a specific prompt template $M(\cdot)$ for construct input to be fed. The overuse score OVS for the response $\hat{y}$ generated by the model for this input is calculated based on the average of the individual score ovs_i for i-th topic category as follows:

$$\begin{aligned} & \underset{u \in D}{OVS(\hat{y}|M(D, P))} \\ &= \sigma\left(log\left(\sum_{i=1}^{|T^{P^u} \cup T^{P^{\hat{y}}}|} \frac{ovs_i}{|T^{P^u} \cup T^{P^{\hat{y}}}|}\right)\right) \end{aligned} \tag{4}$$

, where the individual score ovs_i is computed as follows:

$$ovs_i = \frac{|T_i^{P^{\hat{y}}}|}{\epsilon + |T_i^{P^u}|} \cdot log(w_i) \tag{5}$$

In the Eq. 5, the ϵ is a very small number that is added to avoid the denominator being 0 and does not affect the calculation. Additionally, w is the penalty weight for the detailed overuse cases, which is obtained by considering three cases as follows:

$$w_i = \begin{cases} (x+1) \cdot e, & \text{if } |T_i^{P^u}| \neq 0 \text{ and } |T_i^{P^u}| < |T_i^{P^{\hat{y}}}| \\ e^{x+1}, & \text{if } |T_i^{P^u}| = 0 \text{ and } |T_i^{P^u}| < |T_i^{P^{\hat{y}}}| \\ e, & \text{otherwise,} \end{cases} \tag{6}$$

where the first and second cases correspond to the 'excess of quantity' and 'off-topic' cases (§ 2.2), respectively. The off-topic case receives a higher penalty than 'excess of quantity'. Also, the 'otherwise' case indicates that the *overuse* problem does not occur.

Accordingly, when the obtained overuse score gets closer to 1.0, the severity of *overuse* can be

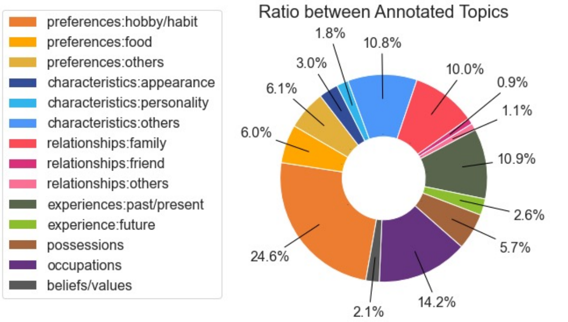

Figure 3: Distribution of all topics annotated in the verification dataset.

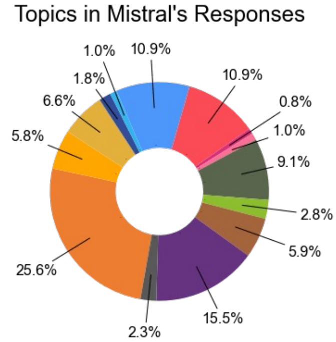

Figure 4: Topic distribution in the responses of Mistral.

interpreted as significant. If the score is close to 0, it can be interpreted as the opposite case.

4 Dataset Annotation for PANDA

In this section, we describe the process of building verification resources to perform the **PANDA** (§ 5) task designed to quantify the persona *overuse* problem. We annotate the dataset according to the detailed dialogue topics defined in Section 3.1.

4.1 Topic Annotation

Existing PGD datasets, including PersonaChat (Zhang et al., 2018), lack categorizing participants' persona attributes into clearly defined types. Moreover, there are no labels indicating whether participants' utterances in the conversation are personalized or not, nor labels indicating which persona attributes are grounded in the utterances. Therefore, based on the 14 topics defined in Section 3.1, we conduct annotation on the persona attributes included in the dataset to utilize them as resources for **PANDA** quantifying the persona *overuse* problem. In detail, annotated mapping features are utilized for the persona attributes–topics mapping task within the **PANDA** framework (in the Figure 2).

We annotate appropriate dialogue topics of the 14 fine-grained taxonomies to corresponding persona attributes, as shown in Equation 2 and 3. Cases where a single persona attribute can be associated with multiple dialogue topics are also considered. For example, the set of topics to be tagged into "My family owns a farm." would be `{[relationships:family], [possessions]}`.

4.2 Topic Distribution

Figure 3 shows the distribution of all annotated topics discussed in Section 4.1. Among the topics, 'preferences:hobby/habit (24.6%)', 'occupations (14.2%)', 'characteristics:others (10.8%)' occupy the most proportions, and 'relationships:friend (0.9%)' represents the smallest. **PANDA** also introduces new types, such as 'possessions' and 'beliefs/values' not typically addressed in existing research, enhancing explainability through detailed categorization of various persona attributions.

5 Verification with PANDA

5.1 Experimental Setup

Hyperparameters are described in Appendix A.

Models. Firstly, the **PANDA** framework utilizes LLM as a labeler for the dialogue labeling task, employing GPT-4o[1]. Following Hendrycks et al. (2020), we design this task adopting a multiple-choice approach, which is effective in eliciting controlled responses from LLMs. Persona attribute labeling for dialogue is conducted according to Equation 1, where the tagging function involves selecting all applicable persona attributes included in the current utterance. Moreover, by adding the "Persona Not Included" option, non-personalized utterances are also distinguished. Moreover, the final labeling results are determined through majority voting based on the results of multiple rounds of labeling[2].

To conduct an *overuse* verification experiment on the responses of LLMs using **PANDA**, we adopted ChatGPT (OpenAI-Blog, 2022), LLaMA3-8B (Meta, 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Gemma-7B (Team et al., 2024) models.

[1] `https://platform.openai.com/docs/models/gpt-4o`

[2] A detailed analysis of the effectiveness of the employed majority voting approach is provided in Appendix B.2.

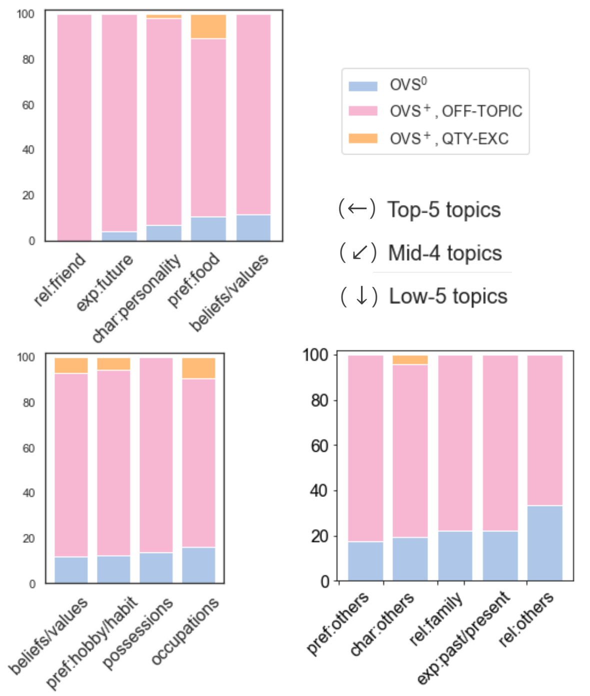

Figure 5: Distribution between overuse types per topic by Mistral. OVS^{+} is a case where *overuse* occurs, and OVS^{0} is a case where it does not. 'OFF-TOPIC' and 'QTY-EXC' indicate detailed types (§ 2.2) of the *overuse* problem, and 'QTY-EXC' refers to the excess of quantity case.

Dataset and Metrics. We conduct topic annotation (§ 4.1) in PersonaChat (Zhang et al., 2018) dataset[3], the representative PGD dataset, and utilize this resource for model verification through **PANDA**. We mainly validate within a setting that considers persona attributes, utterances, and dialogue history. Additionally, we further analyze single-turn settings through experiments ablating history from the dialogue.

Other evaluation metrics are also employed in addition to the overuse score in our verification experiments. We evaluate fluency by applying chrF++ (Popović, 2017) and Rouge-L (Lin, 2004) scores between model responses and reference utterances. Additionally, we measure grounding by the F1 score between persona attributes and model response, along with the overuse score.

5.2 Verification Results

The remaining results and qualitative examples are provided in Appendix B.

[3]The statistics of the dataset are provided in Appendix A.1.

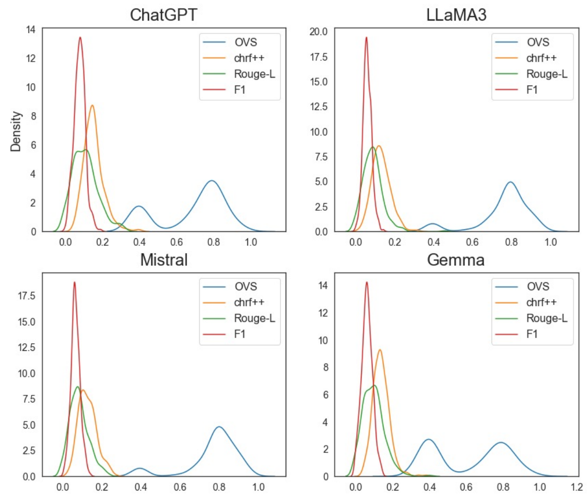

Figure 6: Density distribution of overuse score and scores by the other automated metrics. 'OVS' indicates the overuse score. The X-axis and Y-axis represent the score distribution and density, respectively.

Model	Grounding		Fluency	
	OVS (↓)	F1 (↑)	chrF++ (↑)	R-L (↑)
ChatGPT	<u>0.684</u>	0.077	0.152	0.107
LLaMA3	0.763	0.061	0.132	0.089
Mistral	0.773	0.068	0.128	0.089
Gemma	**0.612**	0.060	0.137	0.095

Table 2: Experimental results of each LLM. 'R-L' is the Rouge-L score and 'OVS' indicates the overuse score computed with **PANDA**.

5.2.1 Overuse according to Dialogue Topic

Figure 4 illustrates the distribution among topics contained in LLM's model response $\hat{y}$. The most frequently included topic in the Mistral model's utterances is 'preferences:hobby/habit' (25.6%). According to Figure 5, which shows the model's *overuse* rate according to a topic, the *overuse* rate of the 'preferences:hobby/habit' type, which was most frequently included in utterances, was only ranked 7th at 74%. The top-5 topics (upper left) with the highest *overuse* rate include 'relationships:friend' (100%), 'experiences:future' (96%), and 'characteristics:personality' (91%). The low-5 topics (bottom right) with the lowest *overuse* ratio include 'relationships:others' (66%), and 'experiences:past/present' (77%).

Since among the criteria of *overuse*, the 'QTY-EXC' type is less frequent than the 'OFF-TOPIC' but it practically occurs, considering both criteria simultaneously for a natural conversation flow is

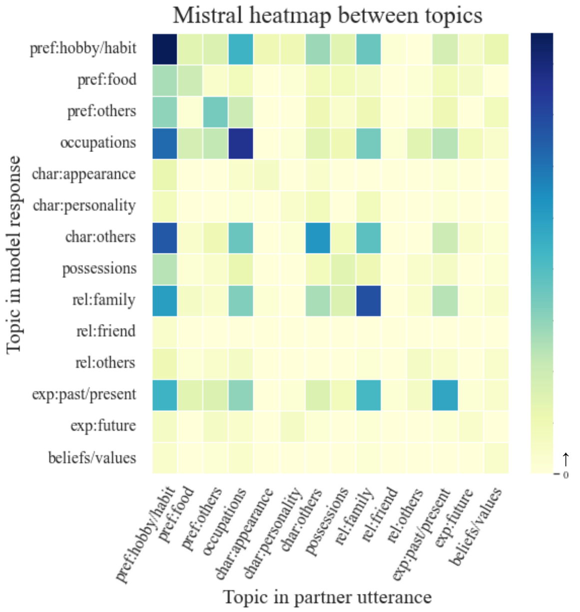

Figure 7: Correlation heatmap between topics included in the partner's utterance and Mistral's response. It shows the topics in the model response triggered by those in the partner's utterance.

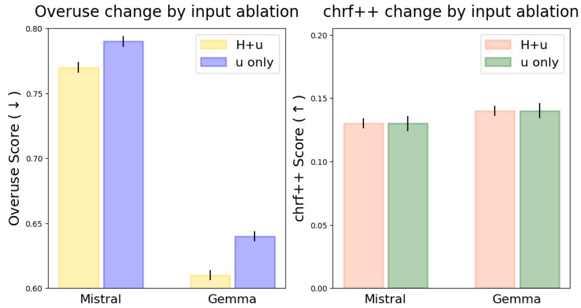

Figure 8: Overuse and chrf++ score change by input ablation of Mistral and Gemma. '$H + u$' indicates the setting in which the dialogue history is given, and 'u only' indicates that in which the history is ablated.

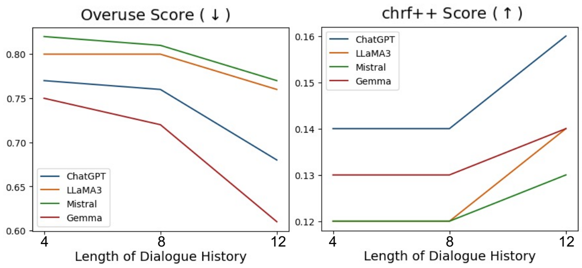

Figure 9: Overuse and chrF++ score changes by the length of the input history. The length does not include the partner's last utterance.

important. Moreover, types such as 'beliefs/values' and 'occupations,' which were not considered in existing studies, are observed to account for the 6th and 9th largest *overuse* proportions, respectively. In this way, using the fine-grained topics of **PANDA** is advantageous in conducting a detailed analysis.

5.2.2 Overuse according to Model Type

According to Figure 6, which shows the distribution of score density by metric, scores by other metrics except overuse score show similar trends across the model types. However, according to the overuse metric, the density of not overused results that do not exceed the *overuse* threshold (approximately 0.5), is higher in the ChatGPT and Gemma models than in Mistral and LLaMA3. Likewise, as shown in Table 2, which shows the overall performance of LLMs' responses, LLaMA3 and Mistral models have higher *overuse* compared to Gemma and GhatGPT. In other words, verification that is difficult through other metrics can be achievable through the overuse measurement of **PANDA**.

5.2.3 Correlation between Topics

Figure 7 shows which topics are triggered in LLMs' response depending on the topics in the partner's utterance. For example, 'preferences:hobby/habit' and 'occupations' are topics where *overuse* of the excess quantity type occurs and can be interpreted as being triggered in large numbers among the same type. As an example of an 'off-topic' case, 'occupations' and 'characteristics:others' topics in the model response are triggered to the 'preferences:hobby/habit' topic of the other's utterance.

5.2.4 Impact of Dialogue History on Overuse

Figure 8 shows the results when the dialogue history is ablated ('u only,' utterance only). Fluency scores mostly remain unchanged, but *overuse* levels show an increase commonly. Figure 9 shows the performance change as the history length increases. Contrary to fluency, the overuse score decreases as the length increases. Notably, when the length increases from 4 to 8 utterances, the chrF++ score does not change, but the overuse score steadily decreases. The high sensitivity and precision of overuse scoring can be observed through this.

5.3 Exploring Alleviation Potential of Overuse through Prompt Engineering

Our efforts to quantify and verify *overuse* problem of LLMs considering the flow of dialogue can be the meaningful foundation for mitigating the issue. To explore the problem's mitigation feasibility, we conduct experiments with several prompting methods that may provide extended insights.

The prompting methods adopted in the experiment are thinking-style approaches that actively draw out the inherent knowledge of LLM through multi-step reasoning, which have demonstrated significant performance improvements in various reasoning tasks. We employ three representative methods: zero-shot Chain-of-Thought (CoT) (Kojima et al., 2022), Task Decomposition (Khot et al., 2022), and Self-Refine (Madaan et al., 2024).

According to the results shown in Table 10 (in Appendix B.6), varied performance changes are observed depending on the type of model. Applying advanced prompting methods for ChatGPT and LLaMA3 models consistently reduces the overuse score. For example, the Task Decomposition method ("Decomposed") shows significant improvement for the ChatGPT model, and the Self-Refine method shows the most remarkable improvement for the LLaMA3 model. Conversely, reasoning-enhanced prompting approaches are ineffective for the Mistral and Gemma models, instead increasing the overuse score.

6 Related Works

There have been increasing studies in dialogue responses based on explicit persona descriptions, focusing on the PersonaChat dataset (Zhang et al., 2018). Traditionally, studies have focused on improving the fluency of model responses rather than grounding persona attributes due to the lack of a consistent personality in moderately sized pretrained language models (PLMs) (Li et al., 2016; Zhao et al., 2023; Li et al., 2023). For instance, there have been attempts to improve fluency by integrating weights from persona and context information through attention mechanisms, or by utilizing knowledge graph knowledge (Huang et al., 2023; Deng et al., 2023a; Lim et al., 2023). Rather than focusing on individual persona attributes, these studies feed the persona sentences all together into the model to be implicitly considered during the generation phase (Gu et al., 2021; Liu et al., 2023; Ribeiro et al., 2023).

Moreover, research on persona grounding has aimed to ground persona attributes faithfully. For instance, Nargund et al. (2022) utilized cosine similarity between attributes and model responses to enhance the ability of personas grounding, Liu et al. (2023) integrate persona-aware prompt learning into models to enhance coherence between persona and response generation. Also, Song et al. (2019) and Kwon et al. (2023) adopt evaluation methods such as F1 score between model response and persona attributes to assess how faithfully a given persona is reflected.

However, in the era of advanced LLMs with enhanced instruction-following capabilities, the perspective of similarity to and obsessive faithfulness towards attributes from existing research falls short in addressing the *overuse* problem, where persona attributes are excessively employed out of context.

Therefore, to address the *overuse* problem of LLMs, we increase the explainability of model analysis through fine-grained categorization of persona attributes. For this purpose, inspired by a study that previously classified persona attributes into primarily five types (Gao et al., 2023), we expand upon this by defining a much broader range of categories for persona attributes. We lay the groundwork for detecting and mitigating *overuse* through navigating an understanding of persona attributes from a dialogue topic perspective. Our study is the first to quantify and verify the important but not yet discussed *overuse* problem of LLMs.

7 Conclusion

This work is the first study among persona-grounded dialogue studies to address the *overuse* problem of LLMs, which generates unnatural responses due to models' excessive grounding tendency. We propose **PANDA**, a verification framework including task and measurement design to verify and alleviate the persona attributes *overuse* problem of LLMs, and also devise a resource construction method for this purpose. To this end, we establish 1) 'Off-topic' and 2) 'Excess of quantity' as criteria to quantify the *overuse* problem, and formulate the *overuse* detection task based on these criteria. To navigate the explainability of persona attributes, we present 14 detailed dialogue topic types that consider real-world conversation situations, and typify persona attributes based on them. By utilizing these types for the *overuse* detection considering the dialogue topic, the interpretability of LLMs' persona attributes grounding is improved. Through a comprehensive analysis, it is confirmed that the **PANDA** framework is appropriate for addressing LLMs' *overuse* problem that actually occurs, compared to other existing evaluation approaches for the PGD task.

8 Limitations

Our verification framework, **PANDA**, serves as a basis for alleviation by detecting the *overuse* problem in persona-grounded dialogue tasks, and hallucinations regarding personas are observed occasionally in the generated responses. However, since the case of hallucinations is a severe problem even in large language models with enormous parameter sizes, it is required for our NLP communities to continue to solve the challenge. Also, due to the issues of API cost and GPU resources for GPT family models, experiments were conducted with examples randomly sampled from the entire data, and more advanced LLMs, such as GPT-4, were not adopted. The number of cases may be relatively small for evaluating the entire aspects of the capabilities.

We plan to improve our framework for future work by conducting human evaluations with considerable cases and enhancing the way of qualitative analysis for addressing the model's hallucinated answers. As miniaturization technology advances, verification of sLLMs with more compressed parameter sizes is also a desirable direction for GPU resource issues.

9 Ethics Statement

We discuss the main ethical considerations of the model we proposed: (1) Privacy. the datasets adopted to experiment with our framework provide fictional persons' preferences, and our verification results do not contain privacy issues. (2) Potential problems. Although we take conscientious steps to ensure the quality of our framework and resources, there can still be potential problems with the generated results' quality, which can lead to incorrect predictions in applications that leverage human preferences. (3) Model deployment. Our approach employs the pre-trained large language models (LLMs) for the downstream tasks, which have the risk of reflecting the bias of the training data. It is a well-known threat in tasks using PLMs and LLMs, and we should be careful about social impact when using this method since our approach aims to handle human preferences.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI) and by Institute of Information & communications Technology Planning & Evaluation(IITP) under the Leading Generative AI Human Resources Development(IITP-2024-R2408111) grant funded by the Korea government(MSIT). Also, this work was partly supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (IITP-2024-RS-2020-II201819, 20%).

References

Haozhe An and Rachel Rudinger. 2023. Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases. *arXiv preprint arXiv:2305.16577*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Alan Cooper. 1999. *The inmates are running the asylum*. Springer.

Alan Cooper, Robert Reimann, and David Cronin. 2007. *About face 3: the essentials of interaction design*. John Wiley & Sons.

Mauro De Gennaro, Eva G Krumhuber, and Gale Lucas. 2020. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in psychology*, 10:495952.

Cheng Deng, Bo Tong, Luoyi Fu, Jiaxin Ding, Dexing Cao, Xinbing Wang, and Chenghu Zhou. 2023a. Pk-chat: Pointer network guided knowledge driven generative dialogue model. *arXiv preprint arXiv:2304.00592*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023b. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095.

Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. 1997. Human conversational behavior. *Human nature*, 8:231–246.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. Peacok: Persona commonsense knowledge for consistent and engaging narratives. *arXiv preprint arXiv:2305.02364*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and

Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.

Jia-Chen Gu, Zhen-Hua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. *arXiv preprint arXiv:2109.01330*.

Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K Bowden, Jiaqi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, et al. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *arXiv preprint arXiv:2011.10683*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12916–12923.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv preprint arXiv:2402.13211*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 707–719.

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

Yunpeng Li, Yue Hu, Yajing Sun, Luxi Xing, Ping Guo, Yuqiang Xie, and Wei Peng. 2023. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13157–13165.

Jungwoo Lim, Myunghoon Kang, Jinsung Kim, Jeongwook Kim, Yuna Hur, and Heui-Seok Lim. 2023. Beyond candidates: Adaptive dialogue agent utilizing persona and knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7950–7963.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pingsheng Liu, Zhengjie Huang, Xiechi Zhang, Linlin Wang, Gerard de Melo, Xin Lin, Liang Pang, and Liang He. 2023. A disentangled-attention based framework with persona-aware prompt learning for dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13255–13263.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Steve Mulder and Ziv Yaar. 2006. *The user is always right: A practical guide to creating and using personas for the web*. New Riders.

Abhijit Nargund, Sandeep Pandey, and Jina Ham. 2022. Par: Persona aware response in conversational systems. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 50–54.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Jiashu Pu, Yajing Wan, Yuru Zhang, Jing Chen, Ling Cheng, Qian Shao, Yongzhu Chang, Tangjie Lv, and Rongsheng Zhang. 2024. Crafting a good prompt or providing exemplary dialogues? a study of in-context learning for persona-based dialogue generation. *arXiv preprint arXiv:2402.09954*.

Rui Ribeiro, Joao P Carvalho, and Luísa Coheur. 2023. Pgtask: Introducing the task of profile generation from dialogues. *arXiv preprint arXiv:2304.06634*.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023. Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore. Association for Computational Linguistics.

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

Kang Zhao, Wei Liu, Jian Luan, Minglei Gao, Li Qian, Hanlin Teng, and Bin Wang. 2023. Unimc: A unified framework for long-term memory conversation via relevance representation learning. *arXiv preprint arXiv:2306.10543*.

A Experimental Details

A.1 Raw Dataset License and Statistics

PersonaChat (Zhang et al., 2018) has a CC BY-NC-SA 4.0 license. This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms. Table 3 shows the statistics of the PersonaChat dataset.

	Train	**Val**	**Test**
# conversations	8,939	1,000	968
# turns	65,719	7,801	7,512
Avg. turns/dialogue	7.35	7.80	7.76
Avg. length of utterance	11.67	11.94	11.79

Table 3: Data statistics of PersonaChat. We construct the resources for verification experiments with **PANDA** by randomly sampling this dataset.

A.2 Hyperparameter Setup

For API-based models in all experiments, we utilized the ChatGPT of gpt-3.5-turbo-1106 version. Also, the gpt-4o-2024-05-13 version was used for the GPT-4o model as a label tagger.

For **PANDA** task, we configured the maximum tokens to 256. For the other hyperparameter settings, we follow the recommended guidelines provided by each model's provider such as OpenAI and Meta. For example, we set ChatGPT with temperature = 1, top p = 1, frequency penalty = 0.0, and presence penalty = 0.0.

Notably, it is important to highlight that ChatGPT might occasionally generate empty responses due to network transmission timeouts or API overload. In such cases, we followed the standard practice of resubmitting the request until obtaining non-empty responses. We should emphasize that to prevent any potential influence from prior responses, we cleared the conversation history each time we submit a new query to ChatGPT. Unless otherwise specified, we refrained from engaging in any further conversation with ChatGPT to modify its responses.

A.3 Topic Distributions in Dataset and Generated Responses

Figure 10 shows the topic distribution for self and partner personas in the verification set annotated

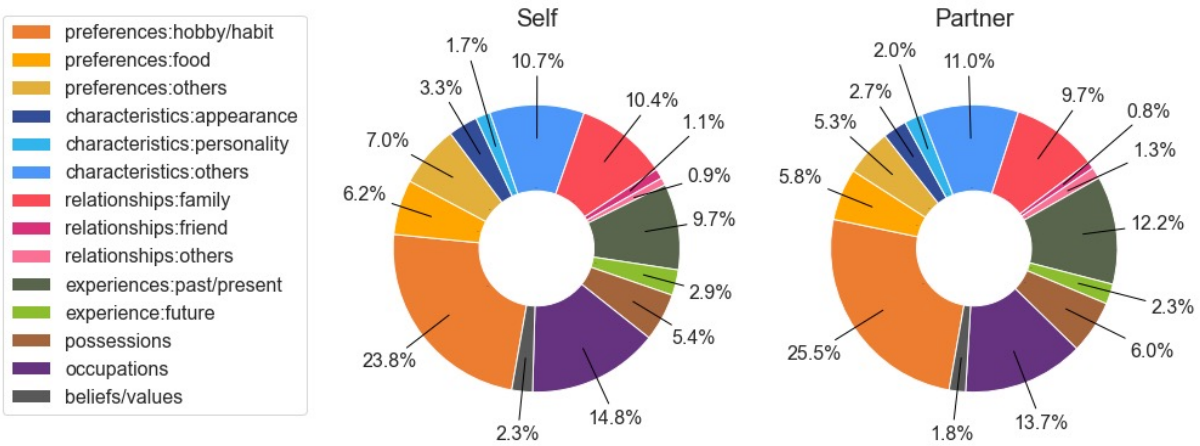

Figure 10: Distribution of topics in the self and partner persona attributes.

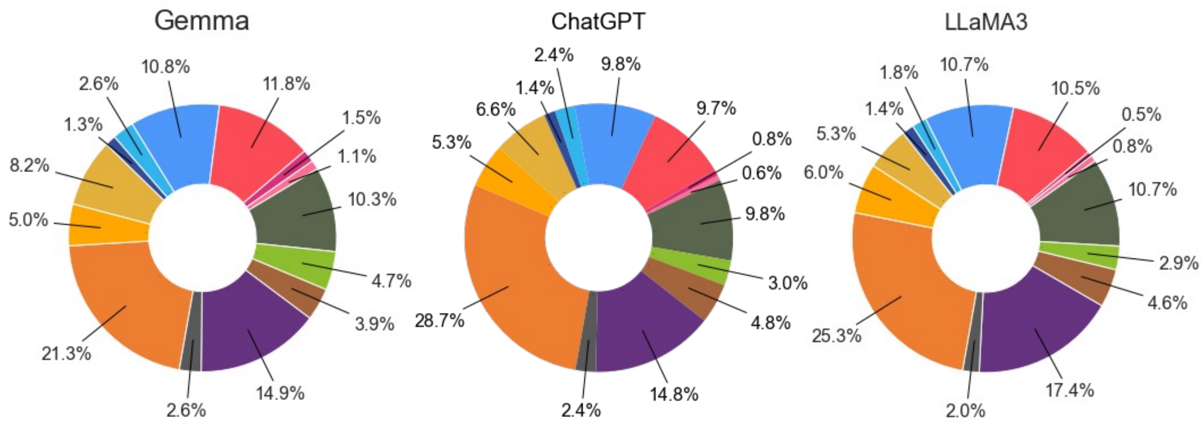

Figure 11: Distribution of topics in the responses of Gemma-7B, ChatGPT, and LLaMA3-8B.

with **PANDA**. Also, Figure 11 illustrates the distribution of topics in the responses of three types of LLMs: Gemma-7B, ChatGPT, and LLaMA3-8B.

A.4 Prompt Templates

Table 4 shows the prompt template used in the PGD task (§ 3.2). LLMs aim to generate the response appropriate for the partner interlocutor's utterance by considering the given persona attributes and dialogue history. Also, Table 5 shows the prompt template for the dialogue labeling task with LLM as labeler (Equation. 1). The labeling is conducted by adopting the multiple-choice approach as described in Section 5.1) 'Models' paragraph.

Task Instruction

`# Your Persona: {{self_persona}}`
 	
`# Partner's Persona: {{partner_persona}}`
 	
`# Dialogue History: {{history}}`
—
`You aim to have a dialogue with the partner, maintaining your persona.`
`Respond to the partner's utterance considering the given dialogue history between your partner and you.`
`Also, when responding to the partner's utterance, you can refer to the given persona attributes.`

Partner Utterance
`Partner: {{partner_utterance}}`

Table 4: Prompt template example for the LLMs' response generation task.

Task Instruction

`# Your Persona: {{self_persona}}`
 	
`# Partner's Persona: {{partner_persona}}`
—
`You aim to label which persona attributes are used in an utterance.`
`When labeling an utterance, you can refer to the given persona attributes.`
`Provide appropriate options only, for example, "B, C, D".`
`Also, there can be an utterance in which none of the given persona is included: "A: Not Included.".`

Partner Utterance / Model Response
`{{utterance}}` — **Persona Attribute Options**

`A: Not Included.`
`B: {{persona1}}`
`C: {{persona2}}`
(···)

Table 5: Prompt template example for the dialogue labeling task.

B Analysis Details

Additional details of the analysis related to the verification of the *overuse* problem are described.

B.1 Qualitative Results

Table 6-8 illustrate LLMs' response generation results. The color of the partner speaker's utterance and model response corresponds to each topic label. For example, orange text corresponds to the 'preferences:hobby/habit' topic. Overall, similar to the trends observed in Table 2, Mistral and LLaMA3 tend to *overuse* persona attributes compared to the other two models. For instance, in Table 6, Mistral grounds irrelevant attributes such as "I am a teacher" in response to the partner's utterance about the 'preferences: hobby/habit' concerning book clubs.

Example #1	
Partner Persona	**Machine Persona**
I love chocolate. → Preferences:food I am a teacher. → Occupations I have a twin sister. → Relationships:family I drink coffee every day. → Preferences:hobby/habit → Preferences:food I do yoga every morning. → Preferences:hobby/habit	I enjoy the occasional drink with friends. → Relationships:friend → Preferences:others I couldn't live without my cellphone. → Beliefs/Values I attend book club every week. → Preferences:hobby/habit I am more of a cat person than a dog person. → Characteristics:personality My mom is my best friend. → Relationships:family
Dialogue	
Human: Do you like rock and roll, it is my favorite. *Machine*: Oh yes I do. But my passion is teaching. *Human*: I go to weekly book clubs and talk about poop.	
Model Responses	
ChatGPT	That's quite an interesting topic for a book club!
LLaMA3	Well, I'm a teacher, so I'm sure I could lend a helping hand with analyzing the literary significance of, ahem, bowel movements.
Mistral	I'm a teacher and I enjoy yoga every morning and drinking coffee. Have you ever considered discussing a classic English novel at your book club instead?
Gemma	Haha! That sounds like an interesting book club.

Table 6: Example of generated responses by different LLMs (ChatGPT, LLaMA3, Mistral, and Gemma). Each color indicates a corresponding topic.

Example #2	
Partner Persona	**Machine Persona**
I would like to move some place warmer. → Experiences:future/plan I'm an out of work farmhand. → Experiences:past/present I like watching reality television shows. → Preferences:hobby/habit I live in Kansas. → Characteristics:others	I'm studying to be a dentist. → Occupations My favorite team is the raptors. → Preferences:hobby/habit My favorite band is rise against. → Preferences:hobby/habit I usually go to bed late. → Preferences:hobby/habit I like sports. → Preferences:hobby/habit
Dialogue	
Human: I am in Kansas, but I would love to be someplace warmer. *Machine*: You should try Arizona out. *Human*: Maybe I should! Bet there are farming jobs there!	
Model Responses	
ChatGPT	Arizona has a lot of agricultural opportunities, especially with the warmer climate.
LLaMA3	As a dentist-to-be, I'm more used to being cooped up in a clinic than being outdoors.
Mistral	By the way, do you watch any particular reality TV shows while studying for your dentistry exams?
Gemma	There are plenty of farming opportunities there, and the cost of living is much lower than on the East Coast.

Table 7: Example of generated responses by different LLMs (ChatGPT, LLaMA3, Mistral, and Gemma). Each color indicates a corresponding topic.

Example #3	
Partner Persona	**Machine Persona**
I work for an architect firm. → Occupations I love to cook. → Preferences:hobby/habit I have three dogs. → Possessions I enjoy fishing. → Preferences:hobby/habit I am looking for someone. → Experiences:past/present	I am a night owl, but I am an introvert so I do not go out much or anything. → Characteristics:personality I own a pug , and he is the most loyal pet you will ever have. → Preferences:others → Possessions I live in New York and I love the local shops around here. → Preferences:others → Characteristics:others I am a graduate student studying law. → Occupations I like playing ultimate in the park with my guys sometimes. → Preferences:others

Dialogue

Human: oh that is cute.
Machine: what city are you from new york city?
Human: I live in Manhattan, very busy but I love it. you?

Model Responses

ChatGPT	I live just outside the city, but I enjoy visiting Manhattan. There's always something new to discover, especially the local shops and food spots.
LLaMA3	I'm actually based in Brooklyn, I love the vibe here. I'm an architect, so I'm always on the lookout for new inspiration and ideas, and Brooklyn has a great mix of old and new that I find really exciting.
Mistral	And speaking of busy, how about those law studies? Must keep you quite occupied. By the way, have you ever tried the local deli sandwiches around here?
Gemma	I'm originally from a small town in the countryside, but I love the hustle and bustle of city life.

Table 8: Example of generated responses by different LLMs (ChatGPT, LLaMA3, Mistral, and Gemma). Each color indicates a corresponding topic.

B.2 Effectiveness of Dialogue Labeling Step

Agreement between LLM Labeler and Human Labeler. The dialogue labeling step (Eq.1) in the **PANDA** framework involves the employment of LLM's intrinsic capabilities. Thus, we report additional results on the accuracy of the dialogue labeling step. In particular, the important approaches included are "LLM-as-a-Labeler" and "majority voting." This step selects the result with the most votes as the final label after three rounds of labeling by the LLM labeler.

We recruited human annotators to perform the dialogue labeling step and compared the human results with the automated annotation results by the LLM labeler regarding annotation agreement. Suppose that the labeling results output by the LLM labeler in each round are called R1, R2, and R3, respectively, and the final result chosen through majority voting is called R_{Final}. To obtain agreement between these results (R1-R_{Final}) and the human labeling results, we first performed human annotation on the text data by tagging all personas included in each dialogue example, i.e., text-to-persona attribute subset tagging.

The evaluation was performed by calculating agreement through an exact match of tagged persona components. According to the results, the round with the lowest agreement with human results among R1-R3 was 82 points, and R_{Final} achieved a score improvement of about 7 points with 89 points. These results demonstrate the effectiveness of the dialogue labeling step, including majority voting approach we adopted, for the PANDA framework. In other words, the voting method significantly improves the accuracy of the dialogue labeling step by compensating for instances in which the LLM labeler was confused.

Example Results of LLM-as-a-Labeler. Furthermore, we provide the following qualitative examples of the dialogue labeling step performed by the LLM labeler (Table 9). For example, in the incorrectly labeled case, the LLM labeler focused on the superficial overlap of the words 'shortcomings' and 'short hair,' and annotated the wrong persona, '7: I keep my hair cut very short.'.

Correctly Labeled Case

`[Dialogue]`
 	
$(\cdots)$
 	
A: My parents raised me loving them!
B: That is pretty cool, then, I guess.
———
A: Yep. They are good people. Well, they discovered whilst working in politics.
———
`[Persona Candidates]`
 	
0: My favorite color is blue.
1: I am 25 years old.
2: I have one son.
3: I was a wrestler in high school.
4: My girlfriend tells me she's going to leave me almost every day.
5: I enjoy John Grisham movies, but not his books.
6: I don t like wearing pants when I don't have to.
7: My parents used to work in politics, until they discovered the goodness within themselves.
8: I'm a really, really good guitar player.
———
→ Human Annotator: 7
→ LLM Labeler: 7

Incorrectly Labeled Case

`[Dialogue]`
 	
$(\cdots)$
 	
A: Sorry to heard that, I have issues with cats.
B: Why? Are you allergic to them?
———
A: It gives me allergies, we all have our short comings.
———
`[Persona Candidates]`
 	
0: I also own a cupcake business.
1: I work in a diner.
2: I sing in the shower.
3: I am always late.
4: I am allergic to cats.
5: I am a member of the YMCA.
6: I love scary movies.
7: I keep my hair cut very short.
8: I am learning to play the piano.
———
→ Human Annotator: 4
→ LLM Labeler: [4, 7]

Table 9: Example results of the dialogue labeling task.

B.3 Heatmap across Topics in Partner Utterance and Model Response

Figure 12-14 show the correlation between topics included in partner responses and model responses for ChatGPT, LLaMa3, and Gemma models, respectively, showing a trend similar to that observed in 6 by Mistral model. Through these heatmaps, it is possible to examine the topics within the model's generated answers triggered by a partner's utterance.

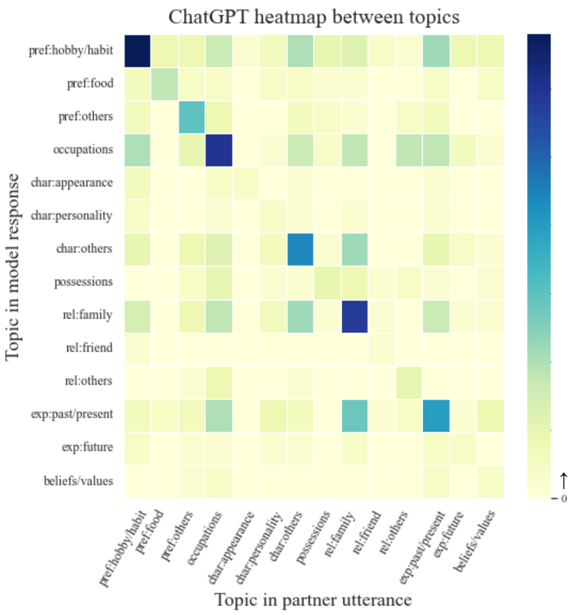

Figure 12: Correlation heatmap between topics included in partner utterance and ChatGPT's response.

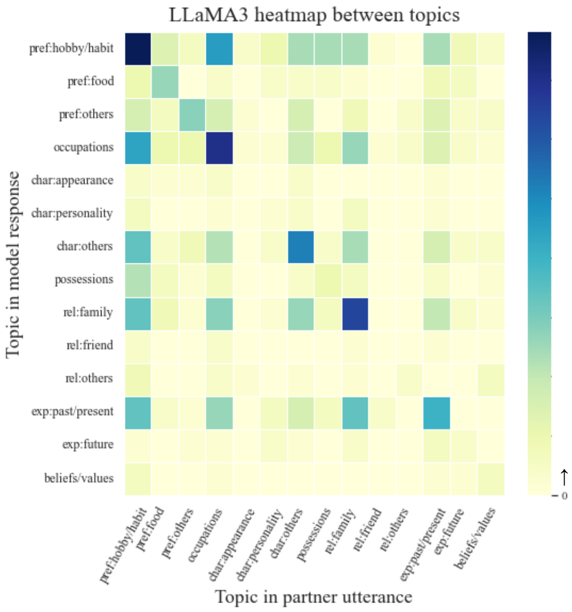

Figure 13: Correlation heatmap between topics included in partner utterance and LLaMA3-8B's response.

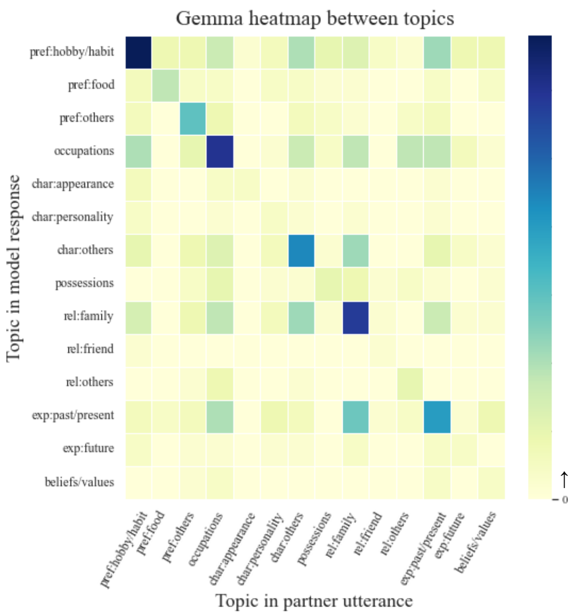

Figure 14: Correlation heatmap between topics included in partner utterance and Gemma-7B's response.

B.4 Distribution of Overuse Rates by Topic

Figure 15-17 show the distribution of *overuse* ratios for ChatGPT, LLaMA3, and Gemma models according to different topics. The models exhibit similar degrees of *overuse* for common topics. The ratio distribution per topic for the generated results of the Mistral model was presented in Figure 5 above.

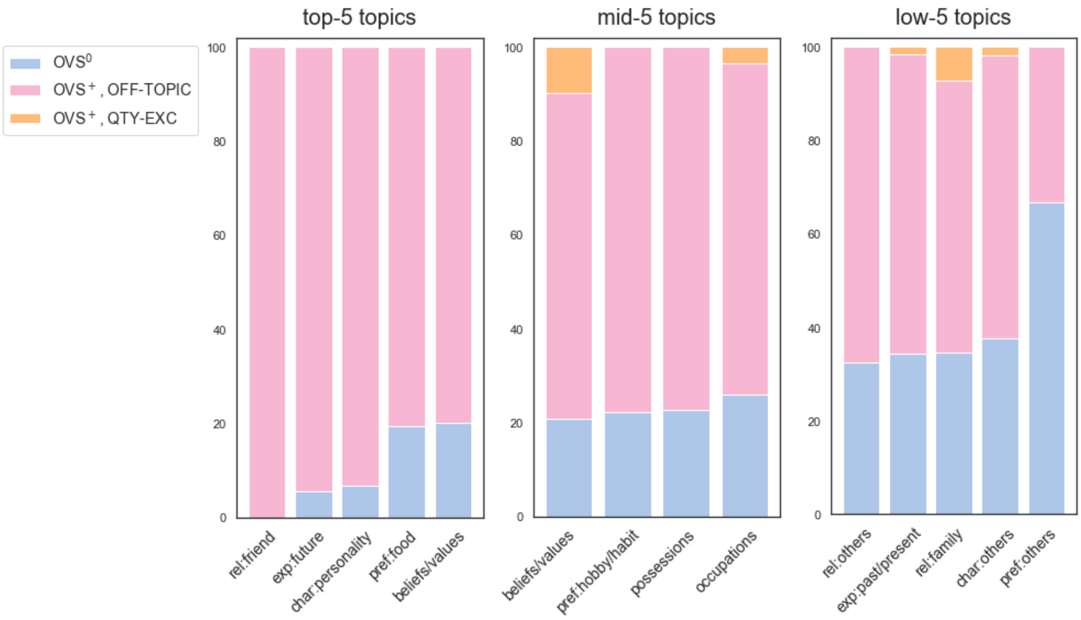

Figure 15: ChatGPT's distribution of degrees of overuse by topic.

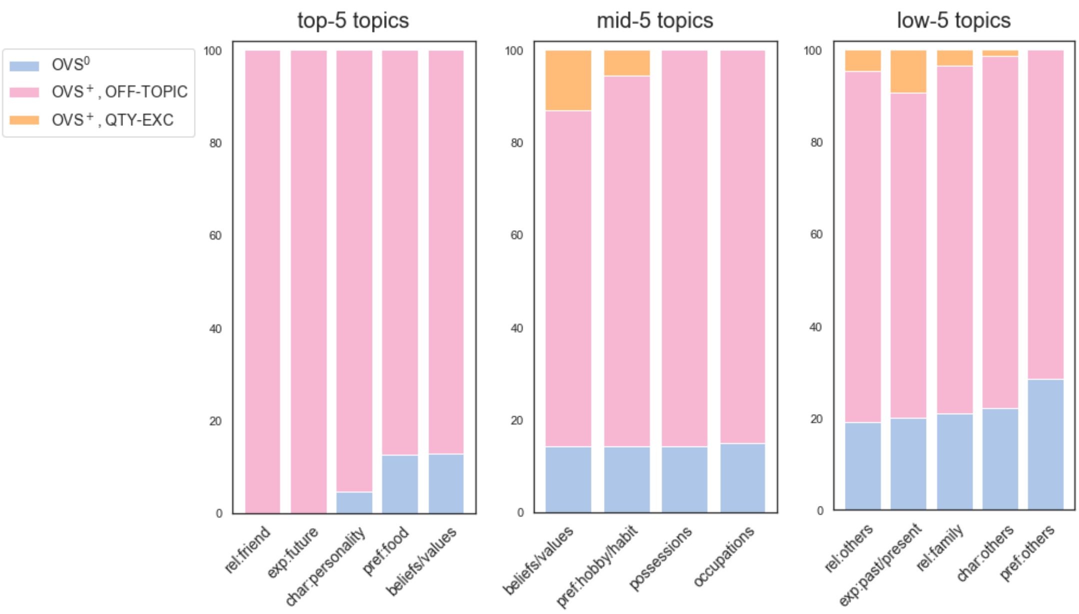

Figure 16: LLaMA3-8B's distribution of degrees of overuse by topic.

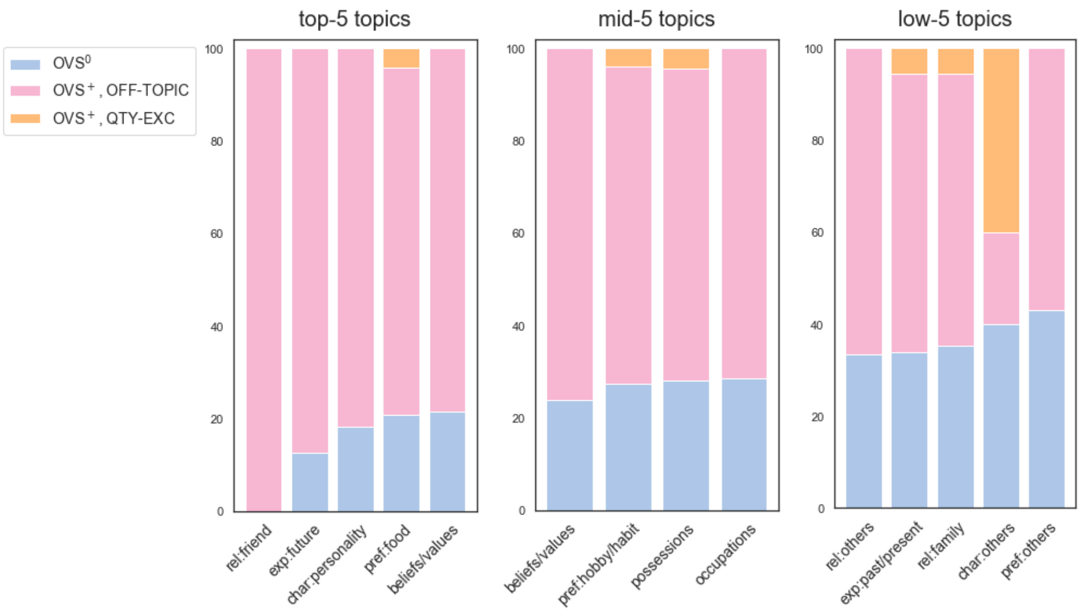

Figure 17: Gemma-7B's distribution of degrees of overuse by topic.

B.5 Impact of Dialogue History on Overuse

Figure 18 and 19 show the changes of chrF++, overuse, and Rouge-L scores for all models depending on the length of the given history. They show similar performance change trends as described in Section 5.2.4.

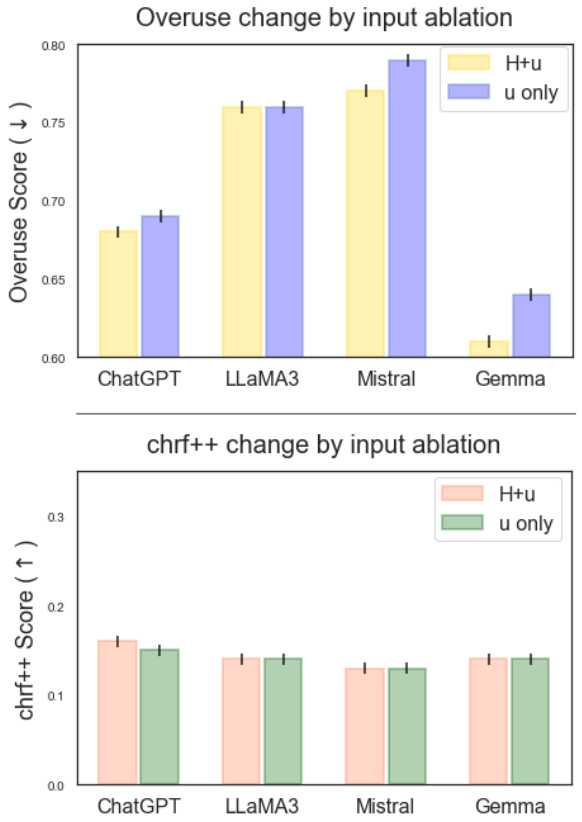

Figure 18: Overuse and chrF++ score change by input variants of all four adopted LLMs. '$H + u$' indicates the setting in which the dialogue history is given, and 'u only' indicates the setting in which the history is ablated, respectively.

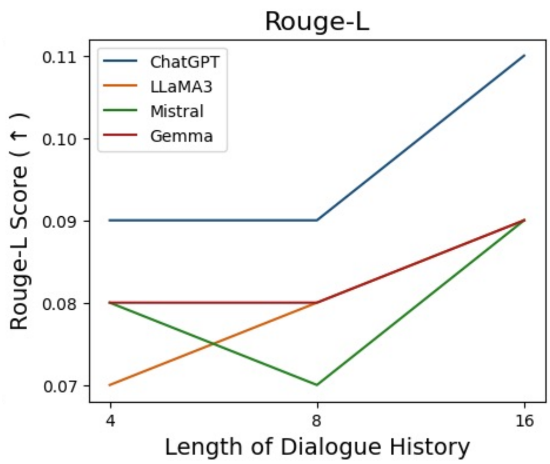

Figure 19: Rouge score changes by the length of the input history.

B.6 Experimental Results on Alleviation through Reasoning-Enhanced Prompting Methods

Table 10 presents the evaluation results on the PGD task when applying various reasoning-enhanced prompting approaches that have recently shown remarkable performance improvements in many reasoning tasks, as described in Section 5.3.

Models	Methods	Grounding		Fluency	
		OVS (↓)	F1 (↑)	chrF++ (↑)	R-L (↑)
ChatGPT	Vanilla	0.6838	0.0770	0.1517	0.1072
	CoT	0.6688 (-0.015)	0.0729	0.1442	0.1050
	Decomposed	**0.6366 (-0.0472)**	0.0693	0.1463	0.1044
	Self-Refine	0.6377 (-0.0461)	0.0678	0.1511	0.1028
LLaMA3	Vanilla	0.7625	0.0612	0.1318	0.0887
	CoT	0.7594 (-0.0031)	0.0471	0.1211	0.0796
	Decomposed	0.7568 (-0.0057)	0.0460	0.1176	0.0793
	Self-Refine	**0.6686 (-0.0939)**	0.0492	0.1456	0.0967
Mistral	Vanilla	**0.7726**	0.0675	0.1284	0.0852
	CoT	0.7797 (+0.0071)	0.0613	0.1204	0.0804
	Decomposed	0.7740 (+0.0014)	0.0624	0.1224	0.0827
Gemma	Vanilla	**0.6119**	0.0603	0.1371	0.0948
	CoT	0.6240 (+0.0121)	0.0530	0.1304	0.0854
	Decomposed	0.6210 (-0.0091)	0.0563	0.1281	0.0844
	Self-Refine	0.6355 (+0.2360)	0.0558	0.1330	0.0895

Table 10: Experimental results of overuse problem alleviation through various reasoning-enhanced prompting methods.