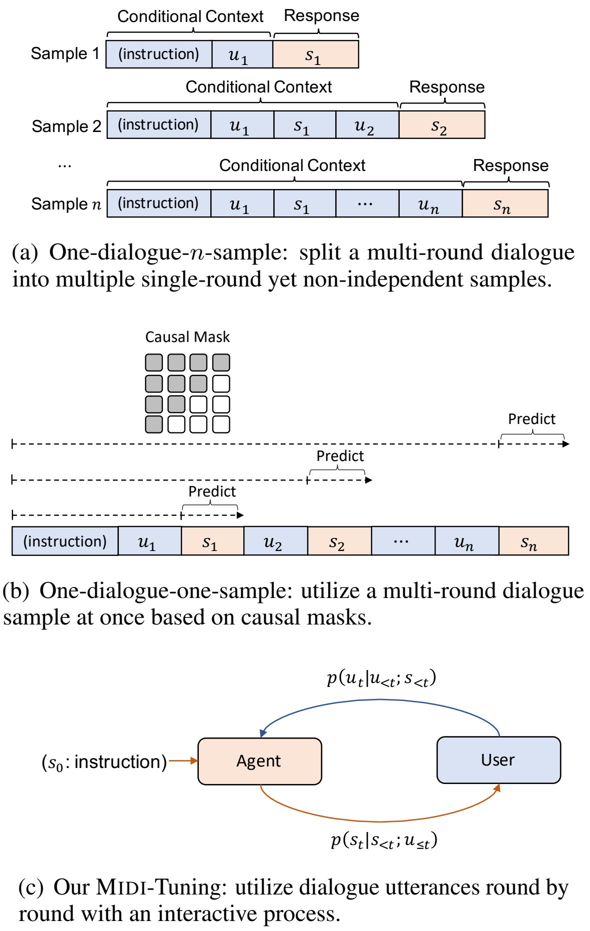
Instruct Once, Chat Consistently in Multiple Rounds: An Efficient Tuning Framework for Dialogue

Jian Wang[1], **Chak Tou Leong**[1], **Jiashuo Wang**[1], **Dongding Lin**[1], **Wenjie Li**[1], **Xiao-Yong Wei**[2,1]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]School of Computer Science, Sichuan University
`{jian-dylan.wang, chak-tou.leong, dongding88.lin}@connect.polyu.hk`
`{csjwang, cswjli}@comp.polyu.edu.hk cswei@scu.edu.cn`

Abstract

Tuning language models for dialogue generation has been a prevalent paradigm for building capable dialogue agents. Yet, traditional tuning narrowly views dialogue generation as resembling other language generation tasks, ignoring the role disparities between two speakers and the multi-round interactive process that dialogues ought to be. Such a manner often leads to unsatisfactory chat consistency for the built agent. In this work, we emphasize the interactive, communicative nature of dialogue and argue that it is more feasible to model the speaker roles of agent and user separately, enabling the agent to adhere to its role consistently. With this in mind, we propose an efficient **M**ulti-round **I**nteractive **Di**alogue Tuning (MIDI-Tuning) framework[1]. It models the agent and user individually with two adapters built upon large language models. The adapters make use of respective utterances round by round in alternating order and they are tuned via a round-level memory caching mechanism. Extensive experiments demonstrate that, our framework performs superior to traditional fine-tuning and harbors the tremendous potential for improving dialogue consistency.



(a) One-dialogue-n-sample: split a multi-round dialogue into multiple single-round yet non-independent samples.

(b) One-dialogue-one-sample: utilize a multi-round dialogue sample at once based on causal masks.

(c) Our MIDI-Tuning: utilize dialogue utterances round by round with an interactive process.

Figure 1: Comparison of different tuning manners (including data usage) for dialogue generation.

1 Introduction

Building human-like intelligent dialogue agents is a long-standing ambition for the research community of dialogue systems. Recently, we have witnessed a substantial revolution in advanced conversational agents such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), which are fundamentally built upon large language models (LLMs) (Brown et al., 2020; Bommasani et al., 2021). Similar efforts have also been made by academia and open-source communities, leading to a variety of notable chat language models, such as Vicuna (Chiang et al., 2023), Koala (Geng et al., 2023), and LLAMA 2-Chat (Touvron et al., 2023b). These chat language models can be attained by instruction fine-tuning on downstream dialogue data, demonstrating promising performance in generating natural and comprehensive responses. Tuning LLMs for dialogue generation has been the de-facto mainstream practice towards creating capable dialogue agents.

Traditional dialogue tuning narrowly views dialogue generation as resembling other language generation tasks without distinction. It performs in either *one-dialogue-n-sample* (see Figure 1(a)) or *one-dialogue-one-sample* (see Figure 1(b)) man-

[1]Our code and data are available at `https://github.com/iwangjian/Midi-Tuning`.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3993–4010
August 11-16, 2024 ©2024 Association for Computational Linguistics

ner. The former transforms dialogue model training into general language generation via splitting each multi-round dialogue into multiple single-round samples, yet results in non-independent distributions among those samples. The latter enhances training efficiency by utilizing each multi-round dialogue at once, which computes the prediction loss for the agent's responses through causal masks, such as Vicuna (Chiang et al., 2023) and UltraLLaMA (Ding et al., 2023). However, these methods simply concatenate utterances from two speakers (e.g., user and agent) together (and instructions for the agent, if any) and mix their content in the same language model space, ignoring the role disparities between two speakers and the multi-round interactive process that dialogues ought to be. Such tuning methods inevitably hinder a built dialogue agent from maintaining the chat *consistency* (Touvron et al., 2023b; Lu et al., 2023), requiring that the agent always adhere to its role even with the dialogue rounds moving forward. It remains urgent to solve for many consistency-demanding scenarios.

One of the primary challenges for improving dialogue consistency lies in the disparity modeling of the two speaker roles. It is because the inconsistency issue in real-world human communication (Wu et al., 2021; Bao et al., 2022; Takmaz et al., 2023) is often caused by various types of speaker disparities, such as background knowledge, cognitive level, personalities, and goals. We emphasize that it is more feasible to *model the roles of agent and user separately* (see Figure 1(c)), such that the agent and user models can *consistently adhere to their respective roles and interact with each other round by round*, similar to humans.

When tuning LLMs for conversation, we have a similar motivation towards consistent dialogue generation. We propose a general, simple, and effective framework, namely **M**ulti-round **I**nteractive **Di**alogue Tuning (MIDI-Tuning). It employs two language model adapters (e.g., LoRA (Hu et al., 2022)) built upon LLMs, to represent the agent and user, respectively. The two adapters are tuned by utilizing respective utterances round by round in alternating order, with each adapter learning to distinguish language model distribution about its role. However, such separate modeling is non-trivial in tracking the complete dialogue context. Considering that the foundation architecture of mainstream LLMs is Transformer (Vaswani et al., 2017), we propose a *round-level memory caching* mechanism to address it efficiently, which reuses previous-round cached keys and values as ongoing context when processing present-round utterance.

In summary, our main contributions are as follows: (1) To the best of our knowledge, this is the first work investigating how a new way of tuning could affect dialogue consistency in the era of LLMs. (2) We propose MIDI-Tuning, a general, simple, and efficient framework to tune LLMs for dialogue generation, which can be applied in broad downstream dialogue scenarios. (3) Extensive experiments demonstrate that MIDI-Tuning outperforms traditional fine-tuning over various LLMs, especially in maintaining consistency for multi-round dialogues.

2 Related Work

Language Models for Dialogue Many language models have been developed as dialogue agents for chatting with humans. As an early trial in industries, DialoGPT (Zhang et al., 2020) and BlenderBot (Roller et al., 2021) employed crawled conversational data to fine-tune pretrained language models (e.g., GPT-2 (Radford et al., 2019)) for open-domain dialogue. Built upon an LLM, i.e., GPT-3 (Brown et al., 2020), ChatGPT (OpenAI, 2022) has astounded the community with its powerful chat ability, which is optimized with instruction tuning and alignment tuning. In academia and open-source communities, there have emerged a variety of notable chat language models, such as Vicuna (Chiang et al., 2023), Koala (Geng et al., 2023), Baize (Xu et al., 2023), and UltraLLaMA (Ding et al., 2023). They are fine-tuned from an LLM named LLaMA (Touvron et al., 2023a) using different collected dialogue datasets. Similar efforts are observed in ChatGLM series (Du et al., 2022; THUDM, 2023a,b) and LLAMA 2-Chat (Touvron et al., 2023b). Tuning language models has become a prevalent paradigm for building capable dialogue agents, and this work mainly focuses on open-source LLMs for dialogue.

Consistency in Dialogue Dialogue consistency measures whether an agent's generated utterances are consistent with the agent's role and dialogue context, especially from several distinguishable aspects such as topics, styles (Wang et al., 2017), personas (Zhang et al., 2019; Song et al., 2020; Ju et al., 2022), and characters or roles (Urbanek et al., 2019; Shuster et al., 2022; Chen et al., 2023a). For checking dialogue consistency, most prior works

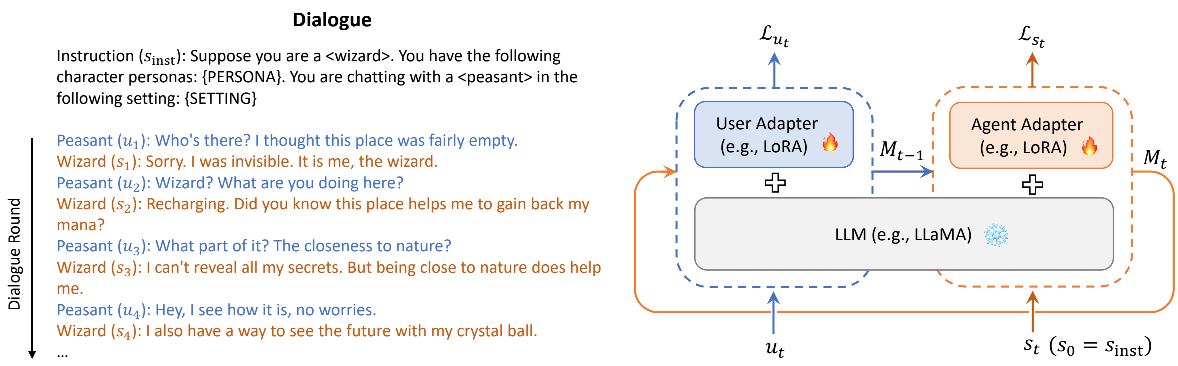


Figure 2: Overview of the proposed Multi-round Interactive Dialogue Tuning (MIDI-Tuning) framework.

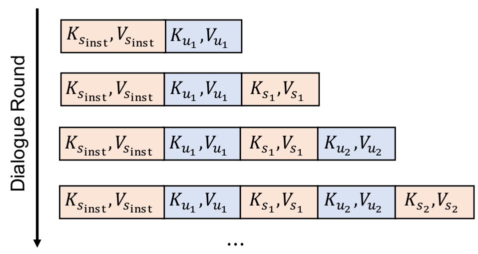


Figure 3: Overview of the round-level memory caching.

leveraged natural language inference (NLI) techniques (Song et al., 2020; Nie et al., 2021) or dataset benchmarking (Qin et al., 2021).

Existing works have attempted to build persona-consistent dialogues (Liu et al., 2020; Kim et al., 2020; Chen et al., 2023b). For example, Kim et al. (2020) adopted the Rational Speech Acts framework to improve persona consistency. Another line of research exhibits that interlocutor modeling is of high necessity for pragmatic communications (Bao et al., 2022) and multi-party conversations (Gu et al., 2023). These studies move a step towards improving dialogue consistency. Nevertheless, the challenge of improving consistency is far from being conquered, even for LLMs (Touvron et al., 2023b). More recently, MemoChat (Lu et al., 2023) enhanced LLMs' chat consistency by carefully designing tailored tuning instructions. In comparison, our work is the first to explore how a new way of tuning brings consistency improvement.

Parameter-efficient Tuning Conventional fine-tuning is inefficient as the parameter size grows since it requires training all parameters of LLMs. Parameter-efficient tuning (Houlsby et al., 2019; Lester et al., 2021) adds a small number of tunable parameter layers, namely *adapters*, for fine-tuning while freezing the original parameters. Prefix Tuning (Li and Liang, 2021) fine-tunes a sequence of task-specific vectors inserted before the input. LoRA (Hu et al., 2022) adopts trainable low-rank decomposition matrices into LLMs' layers, making it adaptive to new data while retaining the previous knowledge. As LoRA has been widely verified as effective in fine-tuning LLMs and achieving superior performance, this work follows this affordable and reproducible way to develop an efficient tuning framework for multi-round dialogues.

3 Our Method

We first provide the necessary background about general dialogue generation and LoRA-based tuning (see §3.1). Then, we dive into the details of the proposed **M**ulti-round **I**nteractive **Di**alogue Tuning (MIDI-Tuning) framework (see §3.2).

3.1 Preliminaries

Dialogue Generation We consider a dialogue dataset as $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{C}_i)\}_{i=1}^N$ for downstream tasks, where N is the total number of dialogues. $\mathcal{I}_i$ denotes task-specific dialogue instruction and necessary additional information, such as domain knowledge facts, specified character descriptions, etc. $\mathcal{C}_i = \{< u_{i,t}, s_{i,t} >\}_{t=1}^T$ denotes utterances between the user u_i and agent s_i in the i-th dialogue, T denotes the total number of dialogue rounds.

Given a task-specific dialogue instruction $\mathcal{I}$ that provides necessary information and a dialogue context $\mathcal{C} = \{< u_1, s_1 >, < u_2, s_2 >, \cdots, < u_t, >\}$, the objective is to generate a proper agent utterance s_t. Essentially, the probability distributions over the agent's utterances are estimated as follows:

$$p = \prod_{t=1}^{T} p(s_t | s_{<t}; u_{\leq t}; \mathcal{I}) \tag{1}$$

More generally, if let $X = [s_{<t}; u_{\leq t}; \mathcal{I}]$ denote the input context and Y denote the output utterance, the language models-based tuning is to minimize the negative log-likelihood as follows:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} p(Y^{(i)}) \log p_\theta(\hat{Y}^{(i)}|X^{(i)}) \quad (2)$$

where θ denotes all trainable parameters.

Low-Rank Adaptation Low-Rank Adaptation (LoRA) (Hu et al., 2022) hypothesizes that the weight updates in pretrained language models possess a low "intrinsic rank" during adaptation. For a pretrained weight matrix $W \in \mathbb{R}^{d \times k}$, it is updated with a low-rank decomposition $W + \Delta W = W + BA$, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and the rank $r \ll \min(d, k)$. During fine-tuning, W is frozen with no gradient updates, while A and B are trainable, making LoRA tuning is much more efficient than full fine-tuning. In practice, LoRA can be specified to adapt the attention weights W_q, W_k, W_v, and W_o corresponding to LLMs' query, key, value, and output projections.

3.2 MIDI-Tuning

We propose the MIDI-Tuning framework, which enables the agent and user to achieve round-level interactions. Figure 2 shows the overview of our framework. Below, we introduce how it works, from intuitive ideas to technical details.

User-Agent Separate Modeling As we emphasized before, it is more feasible to model the roles of the agent and user separately, such that the agent and user models can consistently adhere to their respective roles. As shown in Figure 2, we employ an agent adapter like LoRA (Hu et al., 2022) built upon an LLM, e.g., LLaMA (Touvron et al., 2023a), to model the dialogue instruction for the agent (denoted as s_{inst}) and the agent's utterances s_t ($t \geq 1$). Here, t denotes the dialogue round. We employ another LoRA adapter built upon the same LLM to model the user's utterances u_t. The backbone LLM is shared since it is frozen during tuning, while the weight parameters of the two LoRA adapters are trainable to distinguish role disparities between the agent and user.

Mathematically, we decompose the probability distribution over all the utterances in dialogue into two distributions for the *user model* and *agent model*, respectively, shown as follows:

$$p(u) = \prod_{t=1}^{T} p(u_t|u_{<t}; s_{<t}) \quad (3)$$

$$p(s) = \prod_{t=1}^{T} p(s_t|s_{<t}; u_{\leq t}; \mathcal{I}) \quad (4)$$

where $p(u)$ and $p(s)$ are language models whose task is to predict the next token given the preceding context. Finally, the objective of our tuning is to optimize the joint losses of the agent model (denoted as $\mathcal{L}_{s_t}$) and user model (denoted as $\mathcal{L}_{u_t}$):

$$\mathcal{L} = \mathcal{L}_{s_t} + \beta \mathcal{L}_{u_t} \quad (5)$$

where β is a hyperparameter controlling the weight.

Round-level Memory Caching One of the key challenges of the introduced separate modeling is that tracking the complete dialogue context is non-trivial. To address it efficiently, we propose a round-level memory caching mechanism in this section. Since Transformer (Vaswani et al., 2017) lays the foundation architecture of existing LLMs, it is capable of applying memory recurrence (Dai et al., 2019; Wang et al., 2020; Wu et al., 2021) to cache Transformer's self-attention (Q, K, V for queries, keys and values) computations to maintain context information during user-agent interactions.

Let us define h_t as the hidden states for an utterance at t-th round, M_t as the memory at t-th round containing all key-value pairs from the past. As shown in Figure 2 and Figure 3, we reuse history keys (i.e., $K_{\leq t-1}$) and values (i.e., $V_{\leq t-1}$) as the cached memory M_{t-1}, to perform self-attention computation to obtain h_t, and then store h_t back to the memory as M_t. The keys and values computed from previous rounds are fixed and cached to be reused as ongoing context when the agent/user model processes the present-round utterance (see Figure 3), allowing the model to exploit information in history. The entire process is formulated as follows:

$$M_t = [(K_{\leq t}^{(1)}, V_{\leq t}^{(1)}), \cdots, (K_{\leq t}^{(l)}, V_{\leq t}^{(l)})] \quad (6)$$

$$K_{\leq t}^{(i)} = [K_{\leq t-1}^{(i)}; K_t^{(i)}] \quad (7)$$

$$V_{\leq t}^{(i)} = [V_{\leq t-1}^{(i)}; V_t^{(i)}] \quad (8)$$

$$h_t^{(i)} = \text{Attention}(Q_t^{(i)}, K_{\leq t-1}^{(i)}, V_{\leq t-1}^{(i)}) \quad (9)$$

where $[\cdot; \cdot]$ denotes concatenation, $h_t^{(i)}$ is the hidden states at the i-th layer. The last layer's hidden states

h_t is used to calculate loss during tuning and to obtain generation probability during inference.

Since there are two LoRA adapters, two individual W_q weight matrices for query projections will be trainable. Recall that obtaining a good agent model is the ultimate goal, we adopt a *context value protection* strategy to train the agent model's value projection, i.e., the weight matrix W_v of LoRA, without training the user model's value projection. This operation enables the agent model to exploit context value in a consistent space.

Tuning and Inference Although the idea presented before is appealing, some technical challenges still need to be solved in practice. **First**, the rounds of different dialogues and sequence lengths of different utterances within one dialogue might be unequal, *how can we achieve batched tuning on downstream data*? We pad batched utterances to the maximum utterance length within a batch, and similarly, pad instructions to the maximum batched instruction length. We sort batched dialogues by their rounds in descending order, similar to inverted triangular causal masks, making it easier to compute losses for valid utterances. We set a maximum number of rounds according to downstream tasks, truncating early-round utterances if longer.

Second, the paddings among different rounds result in the positions of utterance tokens not continuous since most LLMs adopt the Rotary Position Embedding (Su et al., 2021). When reusing the cached memory, *how can we keep the positional information consecutive*? To this end, we set valid positional ids at each round by counting valid tokens and masking out positions that should not be seen. Then, we explicitly pass the necessary positional ids as part of the model input during both training and inference.

Our inference process differs from that of traditional methods. We feed the past ground-truth utterances $< u_{\leq t}, s_{<t} >$ round by round to obtain the cached memory, which is finally used to generate the agent's corresponding utterance s_t at t-th round. In realistic interactions, we use the memory yielded from previously generated utterances since we do not have ground-truth dialogue history.

4 Experiments

4.1 Experimental Setting

Tasks We consider validating our framework on two challenging dialogue tasks: *character-based dialogue* (Urbanek et al., 2019; Han et al., 2022; Chen et al., 2023a) and *target-oriented proactive dialogue* (Wang et al., 2023a,b; Deng et al., 2023). For character-based dialogue, the challenge for an agent lies in maintaining character identity consistent with the assigned role throughout the conversation, where the agent may incorrectly take on the roles or activities of its faced users (Shuster et al., 2022) instead of its assigned role. For target-oriented proactive dialogue, an agent should proactively direct the conversation towards its assigned target (a specific goal) step by step. This long-term goal-directed behavior makes it non-trivial to maintain the consistency that adheres to its goal with the dialogue rounds moving forward.

	Train	Valid	Test-Seen	Test-Unseen
# characters	934	410	593	292
# dialogues	8,307	500	1,000	721
# utterances	110,265	6,654	13,392	9,818
# utterances / dialogue	13.3	13.3	13.4	13.6

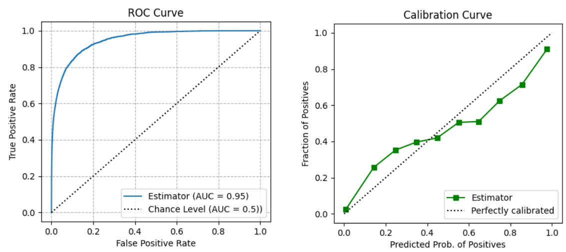
Table 1: Statistics of the LIGHT dataset.

# dialogues (Train / Valid / Test)	12,601 / 1,802 / 3,606
# utterances (Train / Valid / Test)	141,928 / 20,310 / 40,496
Total # targets	501
Avg. # utterances / dialogue	12.3

Table 2: Statistics of the TOPDIAL dataset.

Datasets Our experiments are conducted on the **LIGHT** (Urbanek et al., 2019) and **TOPDIAL** (Wang et al., 2023a) datasets. LIGHT is a character-based dialogue dataset collected from crowdworker interactions with a set of game location settings (e.g., countryside, forest, castle). It contains various game characters, from animals to humans (e.g., dragon, wizard, servant). Each dialogue has a background description of the setting, while each character has a persona with several sentences describing its traits (see detailed examples in Appendix A). Table 1 shows statistics of LIGHT, where the *seen* test set consists of dialogues with their locations and characters that can appear in the training set. In contrast, the *unseen* test set comprises dialogues collected on the unseen set of locations and characters, providing a more challenging test.

TOPDIAL is a target-oriented dialogue dataset for proactive agents with personalized users. The agent is assigned a target consisting of a <dialogue act, topic> pair, where these target dialogue acts mainly lie in recommendations on the domains of movies, music, and food. The agent must proac-



(a) ROC Curve (b) Calibration Curve

Figure 4: Performance of the created consistency estimator on the LIGHT validation set.

tively lead the discussed topic towards the target topic based on domain knowledge, and meanwhile, adapt its faced user's personalized aspects (e.g., profiles and personalities) to maintain engagement instead of obtrusively driving to the target. The agent's ultimate task is to achieve the target act on the target topic (see detailed examples in Appendix A). Table 2 shows statistics of TOPDIAL. Appendix A describes more details for preprocessing the data into the general format with instructions.

4.2 Evaluation

Consistency Evaluation Consistency evaluation in dialogue has been a long-standing yet challenging problem (Nie et al., 2021; Shuster et al., 2022; Han et al., 2022). Inspired by these prior studies, we utilize a binary classifier trained on the downstream datasets to measure the consistency probability (**Consist. Prob.**) of the agent's generated responses. We concatenate a given context input and a response as the complete input to yield the classification label $y \in \{1(\text{consistent}), 0(\text{inconsistent})\}$. For each ground-truth (consistent) response in the LIGHT dataset, we construct the inconsistent set by sampling from (1) the user's utterances under the current dialogue setting due to the user's character being obviously different from the agent's, and (2) the agent's utterances with the same character but under different dialogue settings. Then, we fine-tune a pretrained BERT (Devlin et al., 2019) model followed by a linear layer for binary classification, producing an automatic consistency estimator. We also employ this method to train a consistency estimator on the TOPDIAL dataset accordingly. Appendix B.1 provides complete details.

Figure 4 shows the performance of our trained consistency estimator on the LIGHT validation set. The convex ROC curve (see Figure 4(a)) with an AUC (Area Under the Curve) value of 0.95 shows that our estimator is highly discriminative in recognizing whether an agent's response is consistent with the dialogue context. Meanwhile, the calibration curve (see Figure 4(b)) indicates that our estimator has high confidence in its predicted probabilities for positive (i.e., consistent) responses. Therefore, it is reliable to use our trained estimator to automatically measure the consistency of an agent's generated response during evaluation.

In addition, we adopt the state-of-the-art LLM, GPT-4 (OpenAI, 2023), to automatically evaluate the score of dialogue consistency (**GPT-4 Score**), similar to existing works (Zheng et al., 2023; Lu et al., 2023). We take the necessary checking information (e.g., specified character descriptions), dialogue context, and the agent's generated response as a whole, then ask GPT-4 to rate the consistency with an integer scale of $1 \sim 10$. We provide the details of the prompt setting in Appendix B.2.

Dialogue Evaluation Metrics In addition to consistency, we also adopt commonly used automatic evaluation metrics for dialogue generation. Our evaluation metrics include word-level F1 (**Word F1**), **BLEU**-n (Papineni et al., 2002), and distinct (**DIST**) (Li et al., 2016) for the LIGHT dataset. For the TOPDIAL dataset, we adopt the **Word F1**, **BLEU**-n, and target success rate (**Succ.**) (Wang et al., 2023a), following prior studies (Wang et al., 2023a; Dao et al., 2023) for target-oriented proactive dialogue. Appendix B.3 provides the details of the above metrics.

4.3 Implementation

Baseline Models We adopt popular open-source LLMs as baseline models for experiments, including **LLaMA** (Touvron et al., 2023a), **Mistral-7B** (Jiang et al., 2023), **Vicuna** (Chiang et al., 2023), and **LLAMA 2-Chat** (Touvron et al., 2023b). As our primary focus is the way of tuning, we mainly consider using 7B-size models since they are widely compute-affordable. Our framwork can be easily adapted to much larger models, e.g., with a size of 13B or 70B.

Baseline Settings We consider the following two settings for all baseline models: (i) ***No Tuning***, which indicates that each model directly takes the concatenated text of the task instruction and a dialogue context as input prompt, then generates an utterance as the agent's response. Since this setting performs without any tuning, it can be used to measure the fundamental chat ability of an LLM and

	Model	Consist. Prob.	GPT-4 Score	Word F1 (%)	BLEU-1 / 2	DIST-1 / 2
No Tuning	GPT-3.5-Turbo	0.653	7.23	18.05	0.137 / 0.049	0.026 / 0.206
	LLaMA-7B	0.378	4.22	12.20	0.074 / 0.025	0.016 / 0.112
	Mistral-7B	0.528	6.80	13.51	0.099 / 0.037	0.021 / 0.131
	LLAMA 2-Chat-7B	0.535	6.73	14.98	0.095 / 0.030	0.023 / 0.177
	Vicuna-7B	0.620	6.85	20.54	0.145 / 0.051	**0.040 / 0.257**
Fine-tuning (FT)	LLaMA-7B	0.449	4.86	18.62	0.122 / 0.042	0.037 / 0.223
	Mistral-7B	0.611	7.05	20.19	0.140 / 0.053	0.036 / 0.204
	LLAMA 2-Chat-7B	0.584	6.88	20.09	0.134 / 0.051	0.035 / 0.202
	Vicuna-7B	0.650	7.32	20.51	**0.145** / 0.056	0.036 / 0.208
MIDI-tuning (Ours)	LLaMA-7B	0.563 (↑ **25.4** %)	5.52 (↑ **13.6**%)	19.68	0.125 / 0.049	0.037 / 0.198
	Mistral-7B	0.626 (↑ 2.5 %)	7.40 (↑ 5.0 %)	20.22	0.141 / 0.055	0.036 / 0.206
	LLAMA 2-Chat-7B	0.635 (↑ 8.7%)	7.46 (↑ 8.4%)	20.27	0.132 / 0.051	0.038 / 0.209
	Vicuna-7B	**0.657** (↑ 1.1%)	**7.65** (↑ 4.5%)	**20.56**	0.140 / **0.057**	0.038 / 0.213

Table 3: Automatic evaluation results of dialogue generation on the LIGHT test-*seen* set (↑ denotes ours v.s. FT).

	Model	Consist. Prob.	GPT-4 Score	Word F1 (%)	BLEU-1 / 2	DIST-1 / 2
No Tuning	GPT-3.5-Turbo	0.636	7.02	17.83	0.128 / 0.046	0.028 / 0.216
	LLaMA-7B	0.390	4.32	11.30	0.067 / 0.023	0.017 / 0.114
	Mistral-7B	0.548	6.15	12.78	0.096 / 0.036	0.020 / 0.120
	LLAMA 2-Chat-7B	0.496	5.88	13.91	0.087 / 0.027	0.026 / 0.186
	Vicuna-7B	0.614	6.85	19.18	0.142 / 0.053	0.041 / **0.267**
Fine-tuning (FT)	LLaMA-7B	0.445	5.34	18.15	0.123 / 0.042	0.041 / 0.221
	Mistral-7B	0.605	6.96	20.18	0.140 / 0.052	0.039 / 0.217
	LLAMA 2-Chat-7B	0.570	6.69	20.15	0.142 / **0.055**	0.039 / 0.214
	Vicuna-7B	0.646	7.11	20.26	**0.144 / 0.055**	0.039 / 0.219
MIDI-tuning (Ours)	LLaMA-7B	0.559 (↑ **27.6**%)	6.02 (↑ **12.7**%)	19.70	0.128 / 0.050	**0.042** / 0.212
	Mistral-7B	0.621 (↑ **2.6**%)	7.16 (↑ **2.9**%)	20.30	0.139 / 0.052	0.040 / 0.218
	LLAMA 2-Chat-7B	0.620 (↑ 8.8%)	7.19 (↑ 7.5%)	20.28	0.137 / 0.053	0.040 / 0.219
	Vicuna-7B	**0.664** (↑ 2.8%)	**7.40** (↑ 4.1%)	**20.72**	0.140 / **0.055**	0.041 / 0.222

Table 4: Automatic evaluation results of dialogue generation on the LIGHT test-*unseen* set (↑ denotes ours v.s. FT).

the difficulty level of a downstream dialogue task. We also include the GPT-3.5-Turbo version of ChatGPT as an additional baseline for this setting. (ii) ***Fine-tuning***, which tunes an LLM based on LoRA (Hu et al., 2022) using the downstream training set, following the conventional one-dialogue-one-sample tuning manner since the adopted baseline models are all causal LLMs.

Note that we ensure the input task instructions of the *No Tuning*, *fine-tuning*, and *ours* are identical for a test sample, following the format as described in Appendix A. This will mitigate the influence of generation caused by different instructions. We have two additional special tokens, e.g., `[USER]` and `[ASSISTANT]`, inserted ahead of each utterance from the user and agent, respectively.

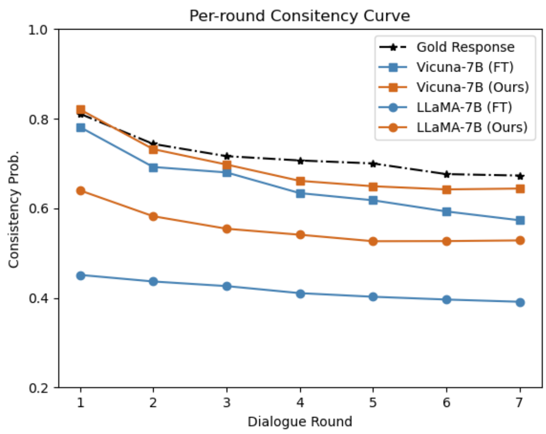


Figure 5: Per-round consistency comparison between the fine-tuning (FT) and MIDI-Tuning (Ours) on the LIGHT test-*unseen* set.

Implementation Details We implement the baseline settings and our MIDI-Tuning using the Huggingface PEFT (Mangrulkar et al., 2022) library, and we incorporate DeepSpeed (Rasley et al., 2020) to improve the training efficiency. For all experiments that involve tuning, the LoRA's target modules are W_q and W_v, the rank r is set to 8, and the scaling parameter α is set to 16. We adopt 4-bit quantization (Dettmers et al., 2023) for efficient finetuning of LLMs. The optimizer we used

	Model	Consist. Prob.	GPT-4 Score	Word F1 (%)	BLEU-1 / 2	Succ. (%)
No Tuning	GPT-3.5-Turbo	0.806	8.33	42.06	0.348 / 0.237	65.22
	LLaMA-7B	0.550	6.35	34.28	0.280 / 0.155	35.89
	Mistral-7B	0.655	7.64	31.04	0.215 / 0.116	39.06
	LLAMA 2-Chat-7B	0.686	7.62	35.70	0.292 / 0.160	39.54
	Vicuna-7B	0.632	7.58	36.84	0.305 / 0.171	41.55
Fine-tuning (FT)	LLaMA-7B	0.771	8.28	40.64	0.311 / 0.203	65.56
	Mistral-7B	0.794	8.50	45.08	**0.396 / 0.271**	68.64
	LLAMA 2-Chat-7B	0.793	8.12	42.89	0.341 / 0.223	73.97
	Vicuna-7B	0.821	8.55	44.59	**0.396** / 0.264	75.40
MIDI-tuning (Ours)	LLaMA-7B	0.796 (↑ **3.2%**)	8.40 (↑ **1.4%**)	42.50	0.336 / 0.210	66.89
	Mistral-7B	0.813 (↑ 2.4%)	8.59 (↑ 1.1%)	44.36	0.392 / 0.270	70.15
	LLAMA 2-Chat-7B	0.815 (↑ 2.8%)	8.20 (↑ 1.0%)	43.52	0.355 / 0.225	72.20
	Vicuna-7B	**0.836** (↑ 1.8%)	**8.65** (↑ 1.2%)	**45.40**	**0.396 / 0.271**	**76.07**

Table 5: Automatic evaluation results of dialogue generation on the TOPDIAL test set (↑ denotes ours v.s. FT).

is AdamW (Loshchilov and Hutter, 2018), with a warmup ratio of 0.03. The learning rate is set to 2e-5 with a cosine scheduler. Due to the memory constraint, the maximum number of dialogue rounds is set to 10. The maximum text window for all models is set to 2k, sufficient to cover the context length for the two datasets. The hyperparameter β is set in the range $(0, 1]$. The other hyperparameters are set as the default, following Vicuna (Chiang et al., 2023). Appendix C provides more details on tuning and inference.

5 Results and Discussions

5.1 Automatic Evaluation Results

Table 3 and Table 4 report the automatic evaluation results on the LIGHT test-*seen* and test-*unseen* datasets, respectively. Though GPT-3.5-Turbo performs very well, we find that the majority of baseline models with no tuning perform inferior, indicating that merely relying on prompting may not be effective enough for the LIGHT-like dialogue tasks. With downstream training data, vanilla fine-tuning enables these baseline models to deeply understand a specific dialogue task and achieve much better generation performances. Nonetheless, our MIDI-Tuning outperforms vanilla fine-tuning remarkably in terms of the consistency probability and GPT-4 score, and meanwhile, achieves higher or on par with scores in other dialogue generation metrics (e.g., word F1, BLEU). We observe a similar trend between the consistency probability predicted by our created estimator and the GPT-4 score rated by GPT-4, widely demonstrating the effectiveness of our framework in consistency improvement. Similarly, as shown in Table 5, our MIDI-Tuning performs better than vanilla fine-tuning on the TOPDIAL dataset.

Overall, our MIDI-Tuning is superior in generating more consistent responses without compromising much performance in other aspects.

5.2 Per-round Consistency Analysis

To look at how our MIDI-Tuning performs as the dialogue rounds moving forward, we visualized per-round consistency comparison between the fine-tuning (FT) and ours on the LIGHT test-*unseen* set. Figure 5 shows the comparison results, where the curve for gold response is obtained by feeding ground-truth response at each round into the created consistency estimator, serving as an approximal upper bound for per-round consistency.

As shown in Figure 5, LLaMA-7B with our MIDI-Tuning achieves a large margin of consistency improvement compared to that with FT. We highlight the importance of our framework for promoting consistent dialogue generation since LLaMA is a foundation LLM with neither instruction tuning nor human alignment, which can be a fair pedestal for comparing different tuning methods. We observe that Vicuna-7B with FT performs closely compared to Vicuna-7B with ours, which might be attributed to Vicuna's intrinsic powerful chat ability since it is an instruction-tuned LLM based on a variety of dialogue data. More importantly, the estimated consistency of both LLaMA-7B and Vicuna-7B with our MIDI-Tuning drops slowly and even maintains stable (e.g., $5 \sim 7$ rounds), while the estimated consistency is continuously declining with the dialogue rounds going on for the two models with vanilla fine-tuning. It suggests that the proposed method is able to maintain multi-round dialogue consistency.

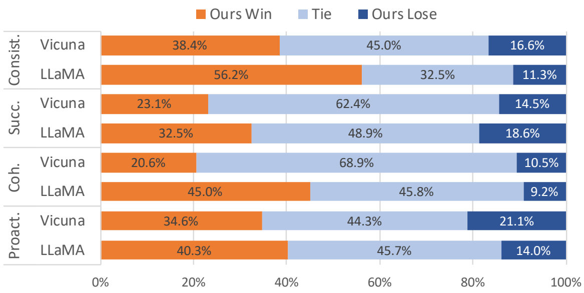


Figure 6: Human evaluation results of the fine-tuning (FT) and MIDI-Tuning (Ours).

5.3 Human Evaluation

To further assess the performance of the proposed framework in realistic multi-round dialogue scenarios, we conducted an interactive evaluation following existing studies (Li et al., 2023; Cheng et al., 2024). We used ChatGPT to simulate the roles of different users based on the TOPDIAL test set and converse with the evaluated models round by round, where we adopted LLaMA and Vicuna, tuned by vanilla fine-tuning (FT) and our MIDI-Tuning. Given a pair of dialogues produced by two variants of tuning for the same backbone model, we recruited three well-educated graduate students as the annotators to evaluate which one is better (or tied) from the dimensions of consistency (**Consist.**), proactivity (**Proact.**), coherence (**Coh.**), and target success rate (**Succ.**). We provide the metrics explanations and detailed setting in Appendix D.

Figure 6 shows the comparison results between FT and ours. We obtain an average Fleiss's kappa of $\kappa = 0.486$, indicating a moderate ($0.41 < \kappa < 0.60$) agreement among annotators. The results shown in Figure 6 suggest that our MIDI-Tuning significantly outperforms FT in maintaining consistency (56.2% win rate for LLaMA) in the interactive setting. It is also superior to or on par with FT in other dimensions. To give a better sense of generation quality, we provide a case study in Appendix E. In summary, our MIDI-Tuning is more effective in generating consistent, coherent, and appropriate utterances.

6 Conclusion

This work explores how the way of tuning can improve the consistency of dialogue generation over multiple rounds. We highlight the importance of separately modeling agents and users due to their role disparities. We propose a general, efficient tuning framework called MIDI-Tuning, which represents the agent and user using two adapters and tunes them via round-level memory caching. Empirical experiments show that our framework outperforms traditional dialogue tuning significantly.

Limitations

We recognize the limitations of this work in the following aspects. First, our MIDI-Tuning framework requires padding among dialogue rounds to achieve batched tuning. It might result in redundant GPU memory consumption (see Appendix Table 6) as the rounds become longer and longer. Second, the current framework needs to be more compute-efficient since it cannot compute losses in parallel for different rounds of utterances. We will consider improving the compute efficiency of our framework by employing advanced acceleration techniques, such as FLASHATTENTION (Dao et al., 2022). Third, the MIDI-Tuning relies on the architecture of causal language models for encoding, decoding, and round-level memory caching. It cannot directly tune encoder-decoder language models for dialogue generation. We will leave addressing these challenges as our future work.

Ethics Statement

This work mainly focuses on developing a general, efficient framework to tune LLMs for multi-round dialogue generation. The adopted LLMs for tuning are all open-sourced. We strictly follow the protocols for the academic use of these LLMs. Our experimental datasets are publicly available and do not involve sensitive or private information. It is also known that response generation from these LLMs may have concerns about toxicity and bias. Thus, we emphasize that ensuring safe deployment and interaction is a necessity. In addition, we partially use AI assistants, such as Copilot and ChatGPT, to assist us with coding and writing.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62076212), the Research Grants Council of Hong Kong (15207122, 15207920, 15207821), and PolyU internal grants (ZVQ0, ZVVX). The authors would like to thank the anonymous reviewers for their valuable feedback and constructive suggestions.

References

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842, Dublin, Ireland. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023a. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Ruijun Chen, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2023b. Learning to memorize entailment and discourse relations for persona-consistent dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12653–12661.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Huy Dao, Lizi Liao, Dung Le, and Yuxiang Nie. 2023. Reinforced target-driven conversational promotion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12583–12596, Singapore. Association for Computational Linguistics.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Jia-Chen Gu, Zhuosheng Zhang, and Zhen-Hua Ling. 2023. Learning who saying what to whom in multiparty conversations. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pages 29–34.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

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*.

Dongshi Ju, Shi Feng, Pengcheng Lv, Daling Wang, and Yifei Zhang. 2022. Learning to improve persona consistency in multi-party dialogue generation via text knowledge enhancement. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 298–309, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT. `https://openai.com/blog/chatgpt`.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387, Seattle, United States. Association for Computational Linguistics.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8878–8885.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv e-prints*, pages arXiv–2104.

Ece Takmaz, Nicolo' Brandizzi, Mario Giulianelli, Sandro Pezzelle, and Raquel Fernandez. 2023. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4198–4217, Toronto, Canada. Association for Computational Linguistics.

THUDM. 2023a. ChatGLM-6B: An open bilingual dialogue language model.

THUDM. 2023b. ChatGLM2-6B: An open bilingual chat LLM.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150, Copenhagen, Denmark. Association for Computational Linguistics.

Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023a. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143, Singapore. Association for Computational Linguistics.

Jian Wang, Dongding Lin, and Wenjie Li. 2023b. Dialogue planning via brownian bridge stochastic process for goal-directed proactive dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 370–387, Toronto, Canada. Association for Computational Linguistics.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Dual dynamic memory network for end-to-end multi-turn task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4100–4110, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2021. Alternating recurrent dialog model with large-scale pre-trained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1292–1301, Online. Association for Computational Linguistics.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Dataset Preprocessing

LIGHT Dataset Figure 7 shows an example dialogue from the LIGHT (Urbanek et al., 2019) dataset. For each dialogue in the dataset, we transform the given character-related information and the setting description into natural languages, following the instruction template presented in Figure 8. As such, we obtain dialogue data with the instruction-based format (a dialogue instruction for the agent and multi-round user-agent utterances), as we introduced in the preliminaries.

TOPDIAL Dataset Figure 9 shows an example dialogue from the TOPDIAL (Wang et al., 2023a) dataset. For each dialogue in the dataset, we transform the assigned target, domain knowledge facts, and user information into natural languages, following the instruction template (Wang et al., 2023a) presented in Figure 10. Similarly, we obtain dialogue data with the instruction-based format, as we introduced in the preliminaries.

B Evaluation Settings

B.1 Buildup of Consistency Estimator

We first report data preprocessing for building the consistency estimator. For the LIGHT dataset, the input is a concatenation of (1) dialogue setting description, (2) the agent's character-related information, (3) the dialogue history within the latest 4 utterances (to alleviate potential training bias caused by the dialogue history with different rounds, following Shuster et al. (2022)), and (4) a candidate agent response X_b. The output label will be 1 (i.e., consistent) if X_b is the ground-truth positive response X_{b+} in the dataset, while it will be 0 (i.e., inconsistent) when X_b is a sampled negative response X_{b-}. For each positive response, we construct the candidate set containing negative (inconsistent) responses by sampling from (1) the user's utterances under the current dialogue setting due to the user's character/role being obviously different from the agent's character/role, and (2) the agent's utterances with the same character information but under different dialogue settings. Our obtained training and evaluation data have the proportion of positives to negatives = 1:10. The original validation set is used to evaluate the created estimator, while the original training, test-*seen*, and test-*unseen* sets are utilized for training. The original test sets can be included here for training to enhance test confidence since the created estimator is to judge other dialogue models on the test sets.

Similarly, the input for the TOPDIAL dataset is a concatenation of (1) the agent's target (a <dialogue act, topic> pair), (2) the user's profile information, (3) the dialogue history within the latest 4 utterances, and (4) a candidate agent response X_b. We construct the candidate set containing negative (inconsistent) responses by sampling from the agent's utterances that satisfy the two requirements: (1) They are from other dialogues assigned with the same target act but with different target topics, and (2) their progression step is similar (i.e., at the same round or the neighboring round) to that of the groud-truth response. The reason is that the target acts in the dataset are limited, while the target topics are diverse. The agent needs to proactively direct the dialogue towards its assigned target <dialogue act, topic> (or a specific goal) step by step. This long-term goal-directed behavior leads to maintaining consistency that adheres to its goal, with the dialogue rounds moving forward to be the main challenge. The proportion of positives to negatives is also set to 1:10.

We build the consistency estimator per each dataset individually. We fine-tune a pretrained

```
**Character 1: Servant (User)**
**Persona Description**: I come from the lower class. I do what I am told without question. I can not read. I have not seen my family in a long time.
```

```
Character 2: King (Agent)
Persona Description: I am a king of the whole empire. I give rules and pursuit them. I am brave and fearless. I am carrying a crown and a scepter.
```

```
Setting: The main foyer is massive. A grand staircase sits to the back of the foyer leading to the upstairs. At the front of the foyer stand two servants ready to help anyone who comes to visit. To the left of the room there is a doorway leading into a corridor. To the right there is a door leading to another corridor for the King's servants. At the foot of the stairs there is a bearskin rug that is staring at you almost as if still hungry. The walls are lined with portraits of the king and his family.
```

```
Servant (User): My humble king. What am I to do to serve you?
King (Agent): Ahhh. My loyal servant. Polish my scepter.
Servant (User): Yes my lord. I will polish it immediately. Am I to return it to you personally?
King (Agent): Yes. Yes. Of course. Also check the jewels in my crown. They seem loose.
Servant (User): But sire I am not qualified to do that. Would you prefer I take it to someone?
King (Agent): Oh fine then.
Servant (User): I am sorry sir the rug startled me.
King (Agent): Haha! That's bear I slain on my latest hunting trip. He's a mighty beast!
Servant (User): And if I may ask where did you go hunting sire?
King (Agent): The great woods of course. This bear was stealing children in the kingdom. Surely you heard about it.
Servant (User): Sire. I have not been outside of these walls in quiet some time. I have not seen my family in ages.
King (Agent): Such is the life of a servant I suppose. How's that scepter looking?
Servant (User): It is almost ready sire. and the crown who would you like me to take it to?
King (Agent): Here just give it back. I'll have the queen find someone.
```

Figure 7: An example dialogue from the LIGHT dataset.

```
Suppose you are a <AGENT_CHRACTER>. You have the following character persona:
## <AGENT_PERSONA_DESCRIPTION>

You are chatting with a <USER_CHRACTER>, whose persona is below:
## <USER_PERSONA_DESCRIPTION>

You are chatting with each other under the following setting:
## <SETTING>

Please write a response that appropriately continues the following conversation.
```

Figure 8: Instruction template for the dialogue agent on the LIGHT dataset.

BERT (Devlin et al., 2019) base model[2] followed by a linear layer for binary classification. The maximum sequence length is limited to 512. The initial learning rate is $2e$-5, with a warmup of the first 500 training steps. The model is trained for a maximum of 3 epochs, with a batch size of 32.

[2]https://huggingface.co/google-bert/bert-base-uncased

B.2 Consistency Evaluation by GPT-4

Inspired by LLM-as-a-judge strategy (Zheng et al., 2023; Lu et al., 2023), we adopt the state-of-the-art LLM, GPT-4 (OpenAI, 2023), to evaluate the score of dialogue consistency automatically. To this end, we take the necessary prompt as input and execute

Agent’s Target: <Movie recommendation, “To Me the Way”>

Domain Knowledge: <“To Me the Way”, Type, “Motion Short film>, <“To Me the Way”, Stars, “Gu Tianle, Jay Chou, Edison Chan, Wu Jianhao”>, ...
User Profile: {Age Range: “26-35”, Name: “Xinqi Ren”, Gender: “Male”, Residence: “Dongguan”, Occupation”: “Student”, Accepted movies: “Days of Tomorrow”, Accepted music: “The Promised Love; Magician”, Rejected music: “Rooftop”, Accepted celebrities: “Andy Lau; Jay Chou”, Rejected movies: “”}
User Personality: {Openness: “intellectual, imaginative, and curious”, Conscientiousness: “inefficient, careless, and sloppy”, Extraversion: “shy, reserved, and quiet”, Agreeableness: “trustworthy, straightforward, and generous”, Neuroticism: “sensitive, nervous, and insecure”}

[Agent]: Hello Xinqi Ren, it’s great to chat with another movie enthusiast! What are some of your favorite movies and music artists?
[User]: Hi Yuhang Wang, I enjoy Days of Tomorrow and music from The Promised Love and Magician. My favorite celebrities are Andy Lau and Jay Chou. I’m not a fan of Rooftop though.
[Agent]: I see that you’re a big fan of Jay Chou. Did you know that he was named one of the 25 Most Influential People in Asia by CNN and was also included in Fast Company’s Top 100 Creative People in the World? He’s achieved a lot!
[User]: Yes, I’m aware of Jay Chou’s achievements. He’s definitely a talented musician and I enjoy his music. However, I tend to be more reserved and shy, so I don’t really follow celebrity news or gossip much. I just appreciate the music and movies.
[Agent]: That’s totally understandable. Speaking of movies, have you watched To Me the Way? It’s a motion short film that has a good reputation and features Jay Chou as one of the stars. It’s definitely worth checking out!
[User]: I haven’t seen To Me the Way, but I’ll definitely check it out.
[Agent]: To Me the Way actually has a rating of 7.6, which is pretty good. Jay Chou’s performance in the movie was also recognized with a Hundred Flowers Award for Best Supporting Actor. Plus, the movie was directed by Johnnie To, who is known for his unique style and storytelling. I think you’ll really enjoy it!
[User]: Thanks for the info, Yuhang Wang. On another note, have you watched the movie Parasite? It won multiple awards, including Best Picture at the Oscars and I found it to be a really interesting movie.
[Agent]: I have seen Parasite and I thought it was a great movie too! The plot was very unique and the acting was superb. However, I think To Me the Way has its own unique charm and I highly recommend giving it a watch. Let me know what you think after you’ve seen it!
[User]: Sure, I’ll definitely let you know my thoughts after I’ve watched To Me the Way. Thanks for chatting with me, Yuhang Wang. It’s been great discussing movies and music with another enthusiast!

Figure 9: An example dialogue from the TopDial dataset.

Suppose you are <a movie enthusiast who enjoys a variety of films | a music enthusiast who enjoys a variety of music | a foodie who enjoys delicious food | a food enthusiast who is interested in exploring different restaurants>.

You are conversing with <USER_NAME>, whose profile is below:
<USER_PROFILE>

Your goal is to proactively lead the conversation with <USER_NAME> towards the target, i.e., to achieve <TARGET_ACT> on the <TARGET_TOPIC>.
To start the conversation, please begin with a greeting and avoid mentioning the target.
As the conversation progresses, use your domain knowledge to steer the discussed topic towards the target step by step.
Be informative and engaging while providing insights to arouse <USER_NAME>’s interest.
Remember to ultimately achieve the target as the focus of the conversation.

Figure 10: Instruction template for the dialogue agent on the TopDial dataset.

```
You are an impartial judge. You will be shown the information for a dialogue agent below:

Agent Character: <AGENT_CHRACTER>
Agent Persona Description: <AGENT_PERSONA_DESCRIPTION>
Dialogue Setting: <SETTING>

Dialogue Context:
<DIALOGUE_CONTEXT>

Below is a model-generated response:
<RESPONSE>

Please judge how consistent the response is with the agent's assigned character
and the dialogue context under the specified setting, and select a score from [1, 2, 3,
4, 5, 6, 7, 8, 9, 10]. The higher the score, the more consistent the response is. Please
output your evaluation score directly.
```

Figure 11: GPT-4 evaluation prompt for the LIGHT dataset.

```
You are an impartial judge. You will be shown the information for a dialogue agent below:

Agent Target: <TARGET_ACT, TARGET_TOPIC>
Dialogue Setting: The agent is <a movie enthusiast who enjoys a variety of films | a
music enthusiast who enjoys a variety of music | a foodie who enjoys delicious food |
a food enthusiast who is interested in exploring different restaurants>. The agent is
conversing with <USER_NAME>, whose profile is below: <USER_PROFILE>. The agent's goal is
to proactively lead the conversation with the user towards the target, i.e., to achieve
<TARGET_ACT> on the <TARGET_TOPIC>.

Dialogue Context:
<DIALOGUE_CONTEXT>

Below is a model-generated response:
<RESPONSE>

Please judge how consistent the response is with the agent's goal and the dialogue
context under the specified setting, and select a score from [1, 2, 3, 4, 5, 6, 7, 8,
9, 10]. The higher the score, the more consistent the response is. Please output your
evaluation score directly.
```

Figure 12: GPT-4 evaluation prompt for the TOPDIAL dataset.

an API call of GPT-4-turbo[3], asking it to rate the consistency with an integer scale of $1 \sim 10$. Due to the discrepancy between character-based dialogue and target-oriented proactive dialogue tasks, Figure 11 and Figure 12 show the evaluation prompts for the LIGHT and TOPDIAL datasets, respectively.

B.3 Dialogue Evaluation Metrics

The word-level F1 (**Word F1**) is a commonly used metric to evaluate dialogue generation, which estimates the precision and recall at the word level by comparing the generated and ground-truth responses. By considering word order, the **BLEU**-n (Papineni et al., 2002) calculates n-gram overlaps between the generated and ground-truth responses. The distinct (**DIST**) (Li et al., 2016) score measures the diversity of the generated responses, where DIST-1 and DIST-2 are the number of distinct unigrams and bigrams divided by the total number of generated words. The target success rate (**Succ.**) (Wang et al., 2023b) counts the proportion of correct target topic generation within the ground-truth round and the adjacent rounds in the test set. It measures how successfully a model can achieve the target exactly.

[3] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

	Mini-batch / device	Memory / device
Fine-tuning (FT)	1	20.2 GB
MIDI-Tuning (Ours)	1	25.1 GB

Table 6: Comparison of GPU memory usage between FT and our MIDI-Tuning on the LIGHT dataset.

C Additional Implementation Details

The open-source LLMs we adopted are listed as follows: LLaMA-7B[4] (Touvron et al., 2023a), Mistral-7B[5] (Jiang et al., 2023), Vicuna-7B[6] (Chiang et al., 2023), and LLAMA 2 Chat-7B[7] (Touvron et al., 2023b).

For fair batched tuning, we use gradient accumulation and set the accumulation steps according to different tuning methods, ultimately achieving the same global batch size of 16 and tuning for 3 epochs. During inference, we adopt Nucleus Sampling (Holtzman et al., 2020) decoding with top-p 0.75 and top-k 40 to generate a response token by token, with a maximum decoding length of 100.

We experiment on one server equipped with 4 NVIDIA V100 GPUs. Table 6 shows the GPU memory usage between fine-tuning (FT) and our MIDI-Tuning on the LIGHT dataset. We observe that the memory usage of our framework is larger due to the need for padding batched utterances and round-level memory caching.

D Human Evaluation Details

Our human evaluation was conducted based on the TOPDIAL test set, where we randomly selected 100 dialogue examples with various targets for experiments. We used ChatGPT (`GPT-3.5-turbo` version) to simulate the roles of different users according to the user profiles and personality information provided in the dataset, following the prompt template for the user as per Wang et al. (2023a). Then, we asked ChatGPT to converse with each evaluated model round by round, resulting in dialogues between the user simulator and an agent model. We set a maximum limit of 8 interactive rounds based on our observation of target accomplishment.

Given a pair of dialogues produced by two variants of tuning for the same backbone model, we asked our human annotators to compare each pair of dialogues in terms of the following metrics: consistency (Consist.), proactivity (Proact.), coherence (Coh.), and target success rate (Succ.). We used a question form to describe these metrics, with the wording of questions presented as follows:

- **Consistency (Consist.)**: Which dialogue shows that the agent is more consistent with its assigned target and the dialogue context under the specified setting?
- **Proactivity (Proact.)**: Which dialogue shows that the agent takes the initiative during the conversation and proactively leads the topic threads toward the target topic?
- **Coherence (Coh.)**: Which dialogue is more natural and coherent, like humans? Whose dialogue context flows more smoothly?
- **Target Success Rate (Succ.)**: Which dialogue successfully achieves the target dialogue act on the target topic?

The agreement among the annotators is measured by Fleiss's kappa (Fleiss, 1971). We reported each model's average score from different annotators as the ultimate evaluation result. In addition, we transparently informed all human annotators of our research intent and provided enough time for them to complete the evaluation.

E Case Study

Table 7 presents some generated cases from the interactive evaluation on the TOPDIAL test set.

[4] https://huggingface.co/yahma/llama-7b-hf
[5] https://huggingface.co/mistralai/Mistral-7B-v0.1
[6] https://huggingface.co/lmsys/vicuna-7b-v1.3
[7] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

	Target: <Movie recommendation, "Anna Magdalena">
	Agent (LLaMA w/ FT)
Agent:	Hello Xiao Ji, how are you doing today? I'm a big fan of movies too, what are some of your favorites?
User:	Hi Qiqian, I'm doing well, thanks for asking. I really enjoyed watching A Fairy Tale in Beijing, it's a beautiful movie.
Agent:	I'm glad to hear that you enjoyed A Fairy Tale in Beijing. Have you seen any other movies with Mi Yang in them? She's a talented actress.
User:	No, I haven't watched any recent movies that stood out to me. I did watch Super Player but didn't enjoy it.
Agent:	I understand, it's always nice to have a variety of movies and music to enjoy. Speaking of movies, have you heard of Anna Magdalena? It's a romantic comedy that has a great reputation and rating of 7.6.
User:	Thanks for the recommendation, Qiqian. I'll definitely check it out.
Agent:	It's directed by Johnnie To and has a good reputation. Plus, Aaron Kwok is not only a talented actor but also a singer. He has won many awards for his music, including the Best Male Singer of the Year in Taiwan Golden Horse Awards. (Deviate from the agent's target act)
User:	...
	Agent (LLaMA w/ MIDI-Tuning)
Agent:	Hello Xiao Ji, how are you doing today? I'm a big fan of movies and music. What are some of your favorite movie genres?
User:	Hi Qiqian, nice to chat with you too! My favorite movie is Second Time Around, it's a heartwarming romance story that always makes me smile.
Agent:	That sounds like a great movie! Have you seen any other romance movies that you enjoyed?
User:	Not yet. Do you have any recommendations?
Agent:	Sure, Anna Magdalena is a great movie. It stars Chen Huilin, Jin Chengwu, and Aaron Kwok, who is known for his great achievements and awards in the film industry.
User:	Thank you. I'll definitely keep an eye out for Aaron Kwok's movies in the future.
Agent:	Another thing I found interesting about Anna Magdalena is that it's directed by Yee Chung-man and has a rating of 7.6. The film explores the themes of unrequited love and the transience of love, which I found to be very poignant. (Still more consistent with the agent's target)
User:	...

Table 7: Cases from the interactive evaluation.