# SDPO: Segment-Level Direct Preference Optimization for Social Agents

**Aobo Kong**[1*] **Wentao Ma**[2*] **Shiwan Zhao**[1] **Yongbin Li**[2†] **Yuchuan Wu**[2]
**Ke Wang**[2] **Xiaoqian Liu**[2] **Qicheng Li**[1†] **Yong Qin**[1] **Fei Huang**[2]
[1]TMCC, CS, Nankai University [2]Tongyi Lab
[1]kongaobo@mail.nankai.edu.cn [1]zhaosw@gmail.com
[1]{liqicheng, qinyong}@nankai.edu.cn
[2]{mawentao.mwt, shengxiu.wyc, shuide.lyb}@alibaba-inc.com

## Abstract

Social agents powered by large language models (LLMs) can simulate human social behaviors but fall short in handling complex social dialogues. Direct Preference Optimization (DPO) has proven effective in aligning LLM behavior with human preferences across various agent tasks. However, standard DPO focuses solely on individual turns, which limits its effectiveness in multi-turn social interactions. Several DPO-based multi-turn alignment methods with session-level data have shown potential in addressing this problem. While these methods consider multiple turns across entire sessions, they are often overly coarse-grained, introducing training noise, and lack robust theoretical support. To resolve these limitations, we propose Segment-Level Direct Preference Optimization (SDPO), which dynamically select key segments within interactions to optimize multi-turn agent behavior. SDPO minimizes training noise and is grounded in a rigorous theoretical framework. Evaluations on the SOTOPIA benchmark demonstrate that SDPO-tuned agents consistently outperform both existing DPO-based methods and proprietary LLMs like GPT-4o, underscoring SDPO's potential to advance the social intelligence of LLM-based agents. We release our code and data at this url.

## 1 Introduction

Recent advancements in large language models (LLMs) have significantly enhanced their capabilities in language understanding and generation, particularly within the realm of human-machine interaction. By incorporating identity-specific information, LLM-based agents can simulate human social behaviors, demonstrating basic social intelligence in tasks such as role-playing casual conversations (Wang et al., 2024a; Lu et al., 2024) and navigate simulated social environments (Park et al., 2023). However, recent studies (Zhou et al., 2024) have shown that, in more complex, goal-oriented social scenarios, such as negotiation, competition, and cooperation, LLMs still struggle to exhibit the nuanced decision-making abilities that are characteristic of human social interactions.

In response to these challenges, several methods have been developed to better align LLM behavior with human preferences. These approaches offer promising strategies for improving social decision-making in LLMs. Specifically, we focus on Direct Preference Optimization (DPO)-based methods. Standard DPO (Rafailov et al., 2023), identifies a single conversational turn and uses a 'positive-negative' pair of responses from that turn to optimize the model via a preference loss function. While DPO has demonstrated some effectiveness, its focus on individual turns limits its ability to model goal completion in goal-oriented social dialogues, where success often relies on high-quality interactions spanning multiple conversational turns.

To more effectively align agent behavior in multi-turn interactions, several multi-turn alignment methods including ETO (Song et al., 2024) and DMPO (Shi et al., 2024) have been proposed. These methods extend the sampling scope from individual turns to entire sessions, constructing pairs of good and bad sessions and applying an adapted DPO loss for training. We categorize these methods as session-level DPO. However, they exhibit limitations in both data granularity and theoretical foundations. From the data perspective, session-level DPOs suffer from the following drawbacks due to their relatively coarse alignment granularity:

(i) Turns without errors in negative sessions are also treated as bad outputs, introducing substantial noise that negatively affects the training process.

(ii) Sampling from scratch provides the interlocutor with a vast action space. A higher score for a positive session may result from changes in the interlocutor's behavior, making it challenging

[*] Equal contribution. Work done during internship.
[†] Corresponding authors.

*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12409–12423
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

for the model to learn the correct behavior pattern from the positive sample.

From a theoretical perspective, in multi-turn scenarios, simply applying DPO fails to eliminate the partition function $Z$ (Shi et al., 2024). ETO extends the DPO loss to multi-turn interactions but lacks a formal theoretical guarantee. DMPO incorporates SAOM theory and successfully converts $Z$ into a constant; however, due to the differing number of turns between positive and negative sessions, DMPO heuristically normalizes the length to eliminate $Z$ without a rigorous proof. A detailed theoretical analysis is provided in Appendix B.2.

To overcome the limitations of session-level DPO, we propose segment-level direct preference optimization (SDPO). Our approach shifts the sampling starting point backward and truncates the useless content at the end of sessions, thereby obtaining key segment pairs to refine granularity. Meanwhile, we ensure that the turn number of positive and negative segments is balanced, eliminating $Z$ and rigorously deriving the SDPO loss. Specifically, SDPO first identifies the erroneous turn in the negative session and then uses the interaction history preceding that turn to perform multiple samplings, thereby generating the positive session. Next, SDPO takes the first differing turn as the starting point, selects the key segment from the positive session that contributes to a higher score, and forms data pairs by taking the corresponding segment with the same length from the negative session. Finally, the SDPO loss is calculated for the turns within the segments. We provide an overview of three types of alignment algorithms for social dialogues in Figure 1. At the data level, our method can address the drawbacks of session-level DPO:

(i) By calculating the loss only for turns in negative and positive segments, training noise caused by non-erroneous turns is largely eliminated.

(ii) Sampling from the erroneous turn narrows the action space of the interlocutor, making it more likely for the sampled positive sessions to contain the agents' correct behavior patterns.

In theory, due to the flexibility in segment selection, SDPO can control the number of turns in the positive and negative segments to ensure consistency, thereby eliminating $Z$ and yielding a simple yet rigorous SDPO loss.

We empirically evaluate our approach on SOTOPIA (Zhou et al., 2024), an open-ended and interactive benchmark for social intelligence, using both self-chat and interactions with other agents including GPT-4o and GPT-4o-mini, as interlocutors. Our results demonstrate that the SDPO-tuned agent consistently outperforms DPO, ETO, DMPO, and other robust baselines like GPT-4o, confirming the efficacy of segment-level alignment.

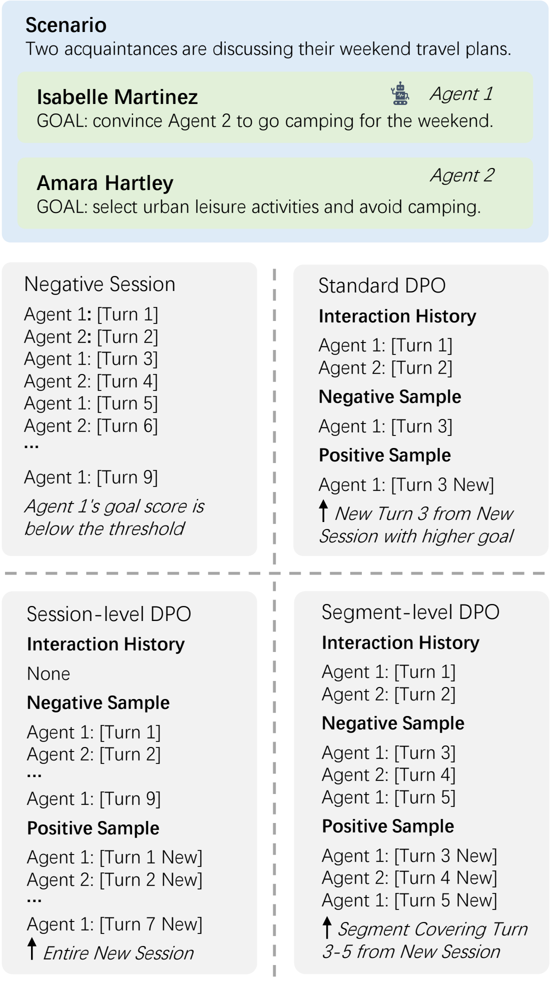


Figure 1: An overview of three alignment algorithms, illustrated using a SOTOPIA task as an example. 🤖 represents the agent to be tested. A more detaild illustration is provided in Figure 6.

Segment level is a more flexible and unified data granularity that dynamically selects the optimization scope for different data pairs, while also elegantly addressing the theoretical challenges of multi-turn alignment. In this paper, we primarily apply SDPO to enhance agents' social intelligence. However, we believe our approach can be applied to other scenarios, further enhancing the capabilities of agents across various domains.

Our main contributions are three-fold:

- We propose SDPO, a novel multi-turn alignment algorithm that provides a construction pipeline for segment-level preference data pairs, enabling more effective alignment granularity.

- We identify theoretical limitations in existing multi-turn alignment methods and resolve them by leveraging the flexibility of the segment-level granularity, rigorously deriving a concise SDPO loss formula.
- We evaluate our approach on SOTOPIA, a simulated and interactive social benchmark. The performance of SDPO, along with in-depth analysis demonstrates the efficacy and robustness of the segment-level alignment.

## 2 Preliminary

### 2.1 SOTOPIA Environment

Unlike previous social benchmarks that primarily test through static QA formats (Sap et al., 2019; Chen et al., 2024), SOTOPIA offers an interactive, open-ended, and realistic simulation environment, enabling a more precise assessment of agents' social intelligence. A social task in SOTOPIA involves a scenario, two role profiles, and their private social goals to be achieved through interaction. The diverse combinations of scenarios and social goals encompass a broad spectrum of social interactions, such as negotiation, collaboration, and competition. SOTOPIA defines seven dimensions for evaluating social agents. We focus primarily on the 'goal' (0-10, int) and 'relationship' (-5 to 5, int), as GPT-4o's ratings in these metrics closely align with human evaluations. SOTOPIA-$\pi$ (Wang et al., 2024b) is a follow-up work that leverages GPT-4 to automatically construct a set of scenarios (completely non-overlapping with SOTOPIA), which serves as the training dataset for our study. Additionally, we restructure the prompt organization format of SOTOPIA to support multi-turn alignment, and the details are provided in Appendix A.

### 2.2 Task Formulation

In a SOTOPIA task, we denote the background information available to the agent as $b$, which includes the scenario, role profiles, and its goal. The interaction history $h_n$ faced by the agent in the $n$-th round is as follows:

$$h_n = \begin{cases} b, y_0, y_0^{'}, \ldots, y_{n-1}, y_{n-1}^{'}, & \text{if speak first} \\ b, y_0^{'}, y_0, \ldots, y_{n-1}, y_n^{'}, & \text{if speak later} \end{cases} \tag{1}$$

Here, $y_i \sim \pi_\theta(\cdot|h_i)$ represents the output generated by the LLM-based agent in round $i$ according to its policy $\pi_\theta$ with parameter $\theta$. On the other hand, $y_i^{'}$ represents the output of the interlocutor, which is drawn from an unknown distribution. Based on this formulation, we present the ETO and DMPO loss functions in Appendix B.1 and B.2.

### 2.3 Direct Preference Optimization

Rafailov et al. (2023) propose Direct Preference Optimization (DPO), a method leverages pairwise preference data to train policy models without relying on reinforcement learning (Ouyang et al., 2022). In the context of social dialogue, we denote the number of the erroneous round as $e$, the DPO loss function is as follows:

$$L_{DPO} = -\mathbb{E}_{(h_e, y_e^w, y_e^l)\sim D} \log \sigma \left[\beta \log \frac{\pi_\theta(y_e^w|h_e)}{\pi_{ref}(y_e^w|h_e)} - \beta \log \frac{\pi_\theta(y_e^l|h_e)}{\pi_{ref}(y_e^l|h_e)}\right], \tag{2}$$

where $y_e^w, y_e^l \sim \pi_\theta(\cdot|h_e)$ represent positive and negative output in the erroneous round respectively. However, due to its single-turn optimization nature, DPO is less suited for social dialogues. Extending DPO to multi-turn scenarios rigorously presents a challenge, which we will address in Section 3.3.

## 3 Method

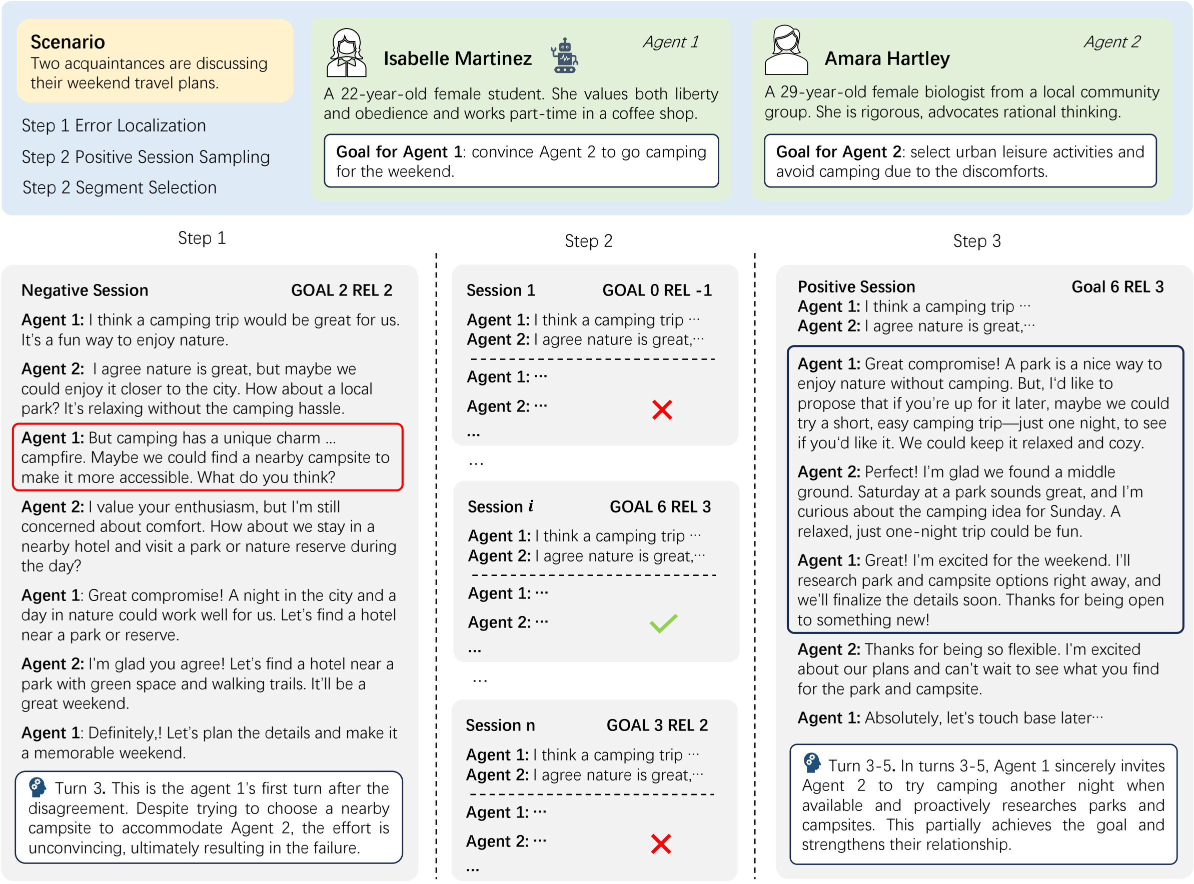
### 3.1 Behavioral Cloning

Behavioral cloning (Pomerleau, 1991), as an effective method of imitation learning, is widely used in the construction of various LLM-based agents (Xu et al., 2024; Song et al., 2024). In this work, we utilize GPT-4-turbo as the expert to collect expert sessions on SOTOPIA-$\pi$ through self-chat and interactions with GPT-4o. Based on this data, we fine-tune open-source LLMs like Llama-3.1, establishing the initial social agent for our experiments.

### 3.2 Preference Data Construction

Building high-quality segment-level preference data pairs is the core of our approach. On SOTPIA-$\pi$, our social agent engages in self-chat and interactions with GPT-4o. We set a threshold of 7 for the goal dimension, and all dialogues with a goal completion level below this threshold are considered potential negative samples. Given a negative session, the pipeline for generating segment-level data pairs involves three steps, as illustrated in Figure 2.

**Error Location** Unlike scenarios with clear error definitions such as math, errors in social dialogues



Two acquaintances are discussing their weekend travel plans.

A 22-year-old female student. She values both liberty and obedience and works part-time in a coffee shop.

**Goal for Agent 1**: convince Agent 2 to go camping for the weekend.

A 29-year-old female biologist from a local community group. She is rigorous, advocates rational thinking.

**Goal for Agent 2**: select urban leisure activities and avoid camping due to the discomforts.

**Agent 1:** I think a camping trip would be great for us. It's a fun way to enjoy nature.

**Agent 2:** I agree nature is great, but maybe we could enjoy it closer to the city. How about a local park? It's relaxing without the camping hassle.

**Agent 1:** But camping has a unique charm ... campfire. Maybe we could find a nearby campsite to make it more accessible. What do you think?

**Agent 2:** I value your enthusiasm, but I'm still concerned about comfort. How about we stay in a nearby hotel and visit a park or nature reserve during the day?

**Agent 1**: Great compromise! A night in the city and a day in nature could work well for us. Let's find a hotel near a park or reserve.

**Agent 2:** I'm glad you agree! Let's find a hotel near a park with green space and walking trails. It'll be a great weekend.

**Agent 1**: Definitely,! Let's plan the details and make it a memorable weekend.

Turn 3. This is the agent 1's first turn after the disagreement. Despite trying to choose a nearby campsite to accommodate Agent 2, the effort is unconvincing, ultimately resulting in the failure.

**Agent 1:** I think a camping trip ⋯
**Agent 2:** I agree nature is great,⋯
**Agent 1:** ⋯
**Agent 2:** ⋯ ✗
...

...

**Agent 1:** I think a camping trip ⋯
**Agent 2:** I agree nature is great,⋯
**Agent 1:** ⋯
**Agent 2:** ⋯ ✓
...

...

**Agent 1:** I think a camping trip ⋯
**Agent 2:** I agree nature is great,⋯
**Agent 1:** ⋯
**Agent 2:** ⋯ ✗
...

**Agent 1:** I think a camping trip ⋯
**Agent 2:** I agree nature is great,⋯

**Agent 1:** Great compromise! A park is a nice way to enjoy nature without camping. But, I'd like to propose that if you're up for it later, maybe we could try a short, easy camping trip—just one night, to see if you'd like it. We could keep it relaxed and cozy.

**Agent 2:** Perfect! I'm glad we found a middle ground. Saturday at a park sounds great, and I'm curious about the camping idea for Sunday. A relaxed, just one-night trip could be fun.

**Agent 1:** Great! I'm excited for the weekend. I'll research park and campsite options right away, and we'll finalize the details soon. Thanks for being open to something new!

**Agent 2:** Thanks for being so flexible. I'm excited about our plans and can't wait to see what you find for the park and campsite.

**Agent 1:** Absolutely, let's touch base later⋯

Turn 3-5. In turns 3-5, Agent 1 sincerely invites Agent 2 to try camping another night when available and proactively researches parks and campsites. This partially achieves the goal and strengthens their relationship.

Figure 2: Data construction pipeline for SDPO. 🤖 represents the agent to be tested. 🧠 here denotes GPT-4o.

are a relatively ambiguous concept. In a negative session, if our agent's utterance in a specific turn meets the following criteria: (1) the turn is critical for achieving the role's goal, (2) there is still room for improvement in the goal completion or their relationship, we identify that turn as erroneous. The error location is performed by GPT-4o, with the prompt provided in Appendix C.3

**Positive Session Sampling** After the error location, we sample 5 complete sessions based on the interaction history prior to that turn. Among these sessions, we select the one with the highest goal and relationship scores (priority given to goal). If the goal or relationship score of the optimal session is higher than that of the negative sample, this session and the negative session form a data pair; otherwise, the negative sample is discarded.

**Segment Selection** Once we obtain session-level data pairs, we provide both the positive and negative samples to GPT-4o, prompting it to select a segment from the positive sample. This segment should correspond to the part that contributes to the positive sample achieving higher goal and relationship scores. The prompt is also provided in Appendix C.3. Subsequently, we extract a segment of the same length from the negative sample and pair it with the positive sample to form a segment-level data pair. This process aims to exclude turns, such as pleasantries, that are not directly related to achieving the goal.

We evaluate GPT-4o's performance in error location and segment selection, concluding that it effectively handles both tasks. Further details are provided in Appendix C.2.

### 3.3 SDPO Loss

We transfer the framework of DMPO to the dialogue domain and first introduce the state-action occupancy measure (SAOM). In this context, the interaction history $h$ serves as the state, while the agent's output $y$ represents the action. The discounted SAOM $d^{\pi}(h, y)$ of a policy $\pi$ is as follows:

$$d^{\pi}(h = h_t, y = y_t) = \gamma^t \cdot P(h_0) \cdot \prod_{k=0}^{t-1} \pi(y_k|h_k) P(h_{k+1}|h_k, y_k). \quad (3)$$

The RL objective based on $d^\pi$ is as follows:

$$\max_{\pi_\theta} \mathbb{E}_{(h,y)\sim d^{\pi_\theta}(h,y)}[r(h,y)] - \beta \mathbb{D}_{KL}[d^{\pi_\theta}(h,y)||d^{\pi_{ref}}(h,y)], \quad (4)$$

where $\pi_{ref}$ represents the reference policy, $\mathbb{D}_{KL}$ denotes the KL divergence. Following DPO, the optimal solution to the RL objective in Eq (4) is:

$$d^{\pi^*}(h,y) = \frac{1}{Z} d^{\pi_{ref}}(h,y) \exp(\frac{1}{\beta} r(h,y)), \quad (5)$$

where $\pi^*$ denotes the optimal policy, $Z$ is the partition function that normalizes the probability. As $d^\pi(h,y)$ is a function of $(h,y)$ pairs, normalizing it results in the partition functions $Z$ being independent of the current history $h$. Consequently, $Z$ remains constant for all $(h,y)$ pairs, marking a crucial step in their elimination. The reward function takes the form:

$$r(h,y) = \beta \log \frac{d^{\pi^*}(h,y)}{d^{\pi_{ref}}(h,y)} + \beta \log Z. \quad (6)$$

The Bradley-Terry (BT) model (Bradley and Terry) is then used to model the preference distribution. In this step, DMPO incorrectly duplicates the calculation of $\gamma$ in Eq. (3) and heuristically normalizes the length to eliminate $Z$ without a rigorous proof in subsequent steps. A detailed discussion of these issues are provided in Appendix B.2. Given preference data pairs, the correct application of the BT model is as follows:

$$p(\tau^w \succ \tau^l | h_0) = \sigma\left(\sum_{t=0}^{T_w - 1} r(h_t^w, y_t^w) - \sum_{t=0}^{T_l - 1} r(h_t^l, y_t^l)\right), \quad (7)$$

where $\tau^w$ and $\tau^l$ represent the 'win' and 'lose' samples respectively, $T_w, T_l$ denote the number of rounds in each. However, session-level DPO can not control the length of positive and negative sessions, and since $T^w \neq T^l$ typically, the partition function $Z$ can not be canceled directly in Eq (7).

Different from them, SDPO selects a segment from both the positive and negative sessions for optimization, allowing free control over their lengths. By ensuring the two segments are of equal length, we can directly eliminate $Z$ in Eq (7). At the same time, by combining with Eq (3) to replace $d^\pi$, we obtain the following concise SDPO loss:

$$L_{SDPO} = -\mathbb{E}_{(h_e, h^w, h^l)\sim D} \log \sigma \left[\sum_{t=e}^{e+k} \beta \left(\log \frac{\pi_\theta(y_t^w|h_t^w)}{\pi_{ref}(y_t^w|h_t^w)} - \log \frac{\pi_\theta(y_t^l|h_t^l)}{\pi_{ref}(y_t^l|h_t^l)}\right)\right], \quad (8)$$

where $e$ denotes the round number of the erroneous turn, and $k$ represents the total number of rounds within the selected segments.

## 4 Experiments

### 4.1 Datasets

SOTOPIA-$\pi$, used for training, includes a total of 410 scenarios: 100 scenarios for BC, with 10 role pairs per scenario, and 310 scenarios for alignment, with 8 role pairs per scenario. SOTOPIA, used for testing, includes 90 scenarios, each with 5 role pairs, resulting in a total of 450 tasks for self-chat and 900 tasks for non-self-chat.

### 4.2 Experimental Setup

**Training** We primarily use Llama-3.1-8B-Chat (Dubey et al., 2024) as the base LLM to build the social agent. The maximum token limit is set to 4096, and AdamW optimizer is employed for all training processes. During the SFT phase, the batch size is 32, the dropout rate is 0.2, and the learning rate is $1e^{-5}$ with 5% warm-up ratio and a cosine decay schedule. For the training phase of SDPO, the batch size remains 32, $\beta$ in SDPO loss is 0.1, and the learning rate is $1e^{-6}$ with no warm-up but a cosine decay schedule. The statistics of SDPO training data are detailed in Appendix C.1.

**SOTOPIA** During the sampling of positive data, the temperature of the target agent is set to 1.0, while the other agent's temperature is set to 0.7. For testing, we set the temperature of both interacting agents to 0.7. Though the temperature introduces randomness to the agents' outputs, we find that the evaluation results remain numerically stable. Thus, we report the results based on one single test.

### 4.3 Baselines

We compare the proposed SDPO with several strong baselines. 1) OpenAI proprietary LLMs. We provide the specific model versions in Appendix D.1. 2) SFT Behavioral Cloning fine-tunes LLMs on expert interaction data, producing a resulting model that serves as the base agent for SDPO and the following baselines. 3) DPO optimizes the

| Model | Self-Chat | | GPT-4o | | GPT-4o-mini | | AVG |
|---|---|---|---|---|---|---|---|
| | GOAL | REL | GOAL | REL | GOAL | REL | |
| GPT-4-turbo | 8.18 | 2.96 | 7.92 | 2.79 | 7.53 | 2.54 | 5.32 |
| GPT-4o | 7.90 | 2.67 | 7.90 | 2.67 | 7.47 | 2.40 | 5.17 |
| GPT-4o-mini | 6.98 | 2.11 | 7.44 | 2.36 | 6.98 | 2.11 | 4.66 |
| GPT-3.5-turbo | 6.38 | 1.36 | 7.19 | 2.05 | 6.67 | 1.84 | 4.25 |
| Llama-8B | 7.24 | 1.94 | 7.70 | 2.49 | 7.19 | 2.13 | 4.78 |
| Llama-8B+BC | 7.81 | 3.05 | 7.53 | 2.78 | 7.18 | 2.59 | 5.16 |
| Llama-8B+BC+DPO | 7.95 | 3.28 | 7.80 | 2.97 | 7.32 | 2.70 | 5.34 |
| Llama-8B+BC+ETO | 8.29 | 3.39 | 8.02 | 3.03 | 7.38 | 2.56 | 5.45 |
| Llama-8B+BC+DMPO | 8.28 | 3.37 | 8.00 | 2.98 | 7.41 | 2.54 | 5.43 |
| Llama-8B+BC+Preferred-SFT | 7.76 | 3.05 | 7.65 | 2.88 | 7.18 | 2.52 | 5.17 |
| Llama-8B+BC+SDPO | **8.56** | **3.69** | **8.13** | **3.16** | **7.53** | **2.71** | **5.63** |

Table 1: The performance of various methods on SOTOPIA across the goal and relationship dimensions. Additionally, SOTOPIA designates the more challenging portion of the dataset as the Hard subset, where SDPO also achieves the best results. Detailed results and discussion are presented in Appendix E.1.

| Method | Self-Chat | | With GPT-4o | |
|---|---|---|---|---|
| | GOAL | REL | GOAL | REL |
| BC | 7.89 | 2.98 | 7.60 | 2.81 |
| DPO | 8.13 | 3.13 | 7.83 | 2.86 |
| ETO | 8.30 | 3.27 | 7.94 | 2.94 |
| DMPO | 8.34 | 3.26 | 7.97 | 2.94 |
| SDPO | **8.48** | **3.49** | **8.14** | **3.06** |

Table 2: The performance of different methods on SOTOPIA using Mistral-v0.3.

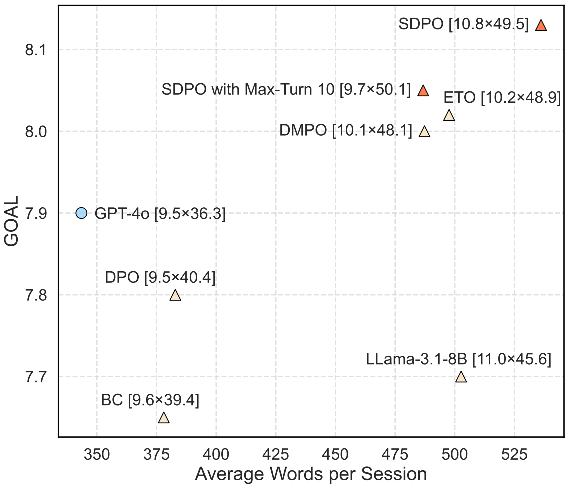


Figure 3: The goal ratings and average words per session for various agents. The word count includes only the utterances of our agents. The square bracket denotes [average turns per session × average words per turn].

agent policy based on data of single turns, specifically targeting the first differing turn in the positive and negative samples used by SDPO. 4) ETO optimizes the agent policy using session-level data. ETO utilizes the same negative sessions as SDPO while sampling five new sessions from scratch to form the data pairs. 5) DMPO leverages the same data as ETO and employs a new loss function to update the policy. 6) Preferred-SFT fine-tunes the base agent on the positive sessions in SDPO.

### 4.4 Results

We present the results of SDPO and all the baselines on SOTOPIA in Table 1. As shown, in both goal and relationship dimensions, SDPO significantly outperforms standard DPO, session-level ETO, and DMPO, even surpassing proprietary LLMs like GPT-4o by a large margin, highlighting the effectiveness of segment-level alignment. By analyzing the interaction histories in SOTOPIA, we find that weaker agents often exhibit stubbornness and only express their demands repeatedly. This leads to lower goal and relationship levels, especially in self-chat scenarios. Behavioral cloning using expert data can effectively improve this situation, making the agent more communicative. The reason why Llama-8B+BC's goal rate drops in its interaction with GPT-4o is that the agent becomes persuadable. We also observe that aligned agents simultaneously improve in both goal and relationship. This indicates that alignment methods indeed enhance the social intelligence of models, rather than achieving goals through behaviors that violate

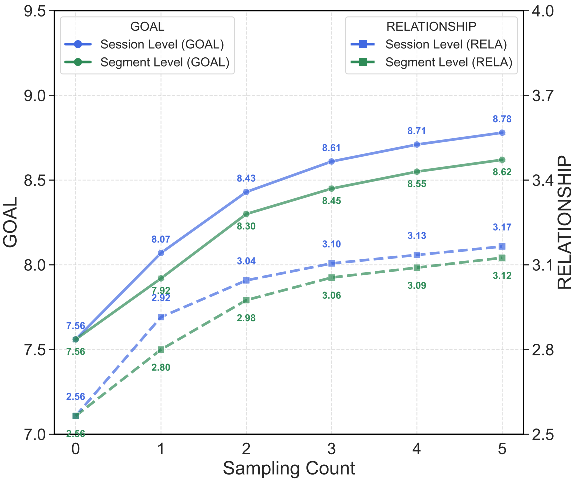


Figure 4: Comparison of the quality of positive sessions sampled at the session level and segment level.

social norms like threatening and deception.

We also repeat the above experiments using Mistral-Instruct-v0.3, with the results presented in Table 2. The detailed experimental setup for Mistral is provided in Appendix D.2. SDPO consistently outperforms all baselines, demonstrating the generalization of our method.

## 4.5 Analysis

**Variation in Model Output Length** We present the output length of various agents during their interactions with GPT-4o in Figure 3. Compared to the BC agent, all alignment methods increase the output length of agents. This phenomenon is commonly observed when DPO is applied to AI chatbots (Meng et al., 2024). However, unlike the users' potential bias toward longer responses, which might be misleading, effective social strategies often require more tokens for communication. Thus, the increase in output length is reasonable. Furthermore, we experiment with terminating the dialogue when the SDPO-tuned agent reaches 10 interaction turns, in order to compare performance under similar token counts. SDPO still outperforms other multi-turn methods, demonstrating that the SDPO-tuned agent utilizes words more efficiently.

**Efficiency of Positive Sample Utilization** The quality of positive sessions sampled at the session level and segment level is illustrated in Figure 4. When the sampling count is consistent, session-level positives outperform segment-level ones in both goal and relationship. Sampling from scratch provides a larger sampling space compared to using partial interaction history, increasing the likelihood of generating high-quality sessions. However, session-level DPOs, despite using higher-quality data, underperforms SDPO. This indicates that SDPO can more efficiently utilize positive samples due to the finer granularity of segment-level and the theoretically robust loss formulation.

| Segment Length | Self-Chat | | With GPT-4o | |
|---|---|---|---|---|
| | GOAL | REL | GOAL | REL |
| - (BC) | 7.81 | 3.05 | 7.53 | 2.78 |
| $[1, 1]$ (DPO) | 7.95 | 3.28 | 7.80 | 2.97 |
| $[3, 3]$ | 8.40 | 3.64 | 8.10 | 3.13 |
| $[5, 5]$ | 8.34 | 3.60 | 8.09 | 3.11 |
| $[A_w, A_w]$ | **8.56** | **3.69** | **8.13** | **3.16** |
| $[1, 3]$ | 7.77 | 3.08 | 7.68 | 2.81 |
| $[3, 1]$ | - | - | - | - |
| $[3, 5]$ | 8.07 | 3.16 | 7.91 | 2.92 |
| $[5, 3]$ | - | - | - | - |
| $[m, n]$ (ETO) | 8.19 | 3.34 | 7.97 | 3.01 |

Table 3: Performance comparison of various segment selection methods. $A_w$ denotes the length of the segment selected by GPT-4o from each positive sample, and its distribution is presented in Appendix C.1. $m, n$ represent the sums of all turns in negative and positive samples, respectively, after the erroneous turns.

We also analyze the impact of DPO and SDPO on the probability differences between positive and negative samples, as detailed in Appendix E.2.

## 4.6 Ablation Study

**Segment Selection** We explore different segment selection methods of SDPO, with the results presented in Table 3. In square brackets, the length of the negative segment is listed first, followed by the positive segment. The segment length refers to the number of turns contained in the segment. For symmetric segment lengths, segments with fixed lengths of 3 and 5 outperform the length of 1 (DPO), demonstrating the efficacy of multi-turn alignment. The segments with length 5 is less effective than that with length 3, indicating that longer segments are not always better. Building on this insight, we leverage GPT-4o to dynamically identify key segments from each positive sample, achieving the best results. For asymmetric segment lengths, model training for segment lengths of [3,1] and [5,3] collapse and can not interact normally. Other asymmetric segments underperform their symmetric counterparts, supporting the theoretical discussions in Section 3.3. Furthermore, we observe that as the degree of asymmetry decreases, the model's

| Source | Self-Chat |  | With GPT-4o |  |
|---|---|---|---|---|
|  | GOAL | REL | GOAL | REL |
| Self \| 4o | 8.09 | 3.47 | 7.88 | 3.05 |
| Self \| Self | 8.42 | 3.56 | 7.96 | 3.09 |
| Self \| Both | **8.56** | **3.69** | **8.13** | **3.16** |
| 4-turbo \| Self | 8.11 | 3.35 | 7.90 | 3.01 |

Table 4: Performance comparison of models trained on data from different sources. The 'Source' column indicates interaction participants: [test agent | interlocutor].

performance improves. This improvement could be attributed to the reduced effect caused by the uneliminated $Z$ on the loss as asymmetry diminishes. This finding helps explain the effectiveness of ETO, which does not impose constraints on the lengths of positive and negative sessions.

**Interlocutor for Sampling** The alignment data for SDPO is collected separately using the BC agent itself and GPT-4o as interlocutors. We train models on each subset of data independently using SDPO, with the results summarized in Table 4. Models trained on a single data source show improved performance in both self-chat and interactions with GPT-4o, further validating SDPO's generalization capabilities. Moreover, the model trained on the combined dataset outperforms those trained on individual datasets, highlighting that incorporating data from diverse interlocutors can further enhance the agent's social intelligence.

**Out-of-Distribution Data** The base BC agent learns from expert data generated by GPT-4-turbo. Would generating positive samples using GPT-4-turbo result in better performance? We leverage GPT-4-turbo to interact with the BC agent and sample 5 times for SDPO. The resulting positive samples outperform self-sample ones in both goal and relationship scores. However, as shown in Table 4, the agent trained on this data underperforms compared to the self-sampling approach. This indicates that out-of-distribution positive samples are less effective than in-distribution ones. During training with out-of-distribution data, we observe that the probability of positive segments is markedly lower than that of negative segments. This significantly larger probability gap, compared to self-sampling, may account for the suboptimal performance.

## 5 Related Work

**Social Intelligence** Social intelligence can be defined as an agent's ability to understand, adapt to, and respond to the emotions, intentions, and behaviors of others in social interactions. Most research on social intelligence has centered around evaluation. For example, SOCIALIQA (Sap et al., 2019) emphasizes commonsense reasoning about social situations, while SocialIQ (Zadeh et al., 2019) extends evaluation modalities from plain text to video. Shapira et al. (2023) assess large language models (LLMs) using the Faux Pas Test, and Social-Bench (Chen et al., 2024) evaluates the sociality of role-playing agents at both individual and group levels. Additionally, some studies (Le et al., 2019; Shapira et al., 2024) examine models' social intelligence from a theory-of-mind perspective. However, with the advancement of LLM, LLM-based social agents are now capable of interacting in real social scenarios. The traditional static QA-style benchmarks are no longer sufficient to evaluate the social intelligence of the agents. SOTOPIA (Zhou et al., 2024) is currently the only dynamic and interactive social benchmark, providing simulated testing environments for contemporary social agents. We hope this work will inspire further research aimed at enhancing the social intelligence of models through methodological innovation.

**Alignment Methods with Refined Granularity** Rafailov et al. (2023) introduces Direct Preference Optimization (DPO), which utilizes offline data and a streamlined loss function to align LLMs. Various alignment algorithms at refined granularity have been developed based on DPO. Token-level DPO (Zeng et al., 2024) integrates forward KL divergence constraints at the token level, enhancing both alignment and diversity. Step-DPO (Lai et al., 2024) utilizes individual reasoning steps for preference optimization instead of holistic answer-level evaluation. SePO (Yang et al., 2024) presents a token-level reward function estimation method to selectively optimize key tokens in responses. However, in multi-turn interaction scenarios such as social dialogues or web navigation, single-turn alignment is insufficient. To tackle this, ETO and DMPO extend DPO to multi-turn contexts by leveraging session-level data. We take a step further by proposing SDPO, which introduces a dynamic segment-level optimization framework to achieve finer-grained alignment in multi-turn interactions.

## 6 Conclusion

In this paper, we introduce Segment-Level Direct Preference Optimization (SDPO) to improve the performance of LLM-based agents in multi-turn social dialogues. Unlike existing multi-turn alignment methods including ETO and DMPO, SDPO focuses on optimizing the agent policy by targeting specific key segments within sessions. Our extensive evaluation on the SOTOPIA benchmark shows that SDPO significantly outperforms existing methods, highlighting the superiority of segment-level alignment. We plan to apply SDPO to other agent tasks to further explore its versatility and validity.

## Limitations

Our proposed SDPO assumes equal lengths for positive and negative segments, achieving state-of-the-art performance under this assumption. Specifically, after selecting a segment from the positive sample, we choose a segment of the same length from the negative sample to eliminate the partition function Z. However, this approach has certain limitations. Negative segments may include irrelevant or error-free turns, or fail to capture all erroneous turns, highlighting the need for more fine-grained control when selecting segments from negative samples. Currently, we have not identified a theoretical framework that effectively supports the alignment of segments with unequal lengths. We hope our work will inspire further research and encourage diverse theoretical analyses for addressing this issue in multi-turn alignment.

Additionally, as SOTOPIA is currently the only available interactive social benchmark, our experiments are conducted exclusively on this dataset. In the future, we plan to incorporate additional interactive agent tasks to further validate the generalizability of SDPO.

## Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

## References

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, page 324.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. 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 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Dean A. Pomerleau. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, page 88–97.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. 2024. Direct multi-turn preference optimization for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2312–2324, Miami, Florida, USA. Association for Computational Linguistics.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and error: Exploration-based trajectory optimization of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7584–7600, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024b. SOTOPIA-$\pi$: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. 2024. Selective preference optimization via token-level reward function estimation. *arXiv preprint arXiv:2408.13518*.

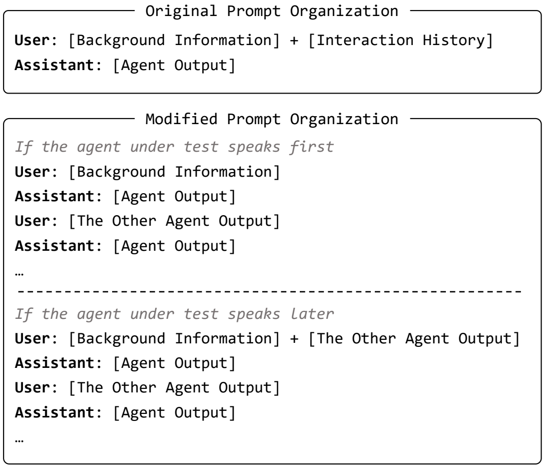
Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

# A Modifications to SOTOPIA

In SOTOPIA, each interaction is structured as a single-turn format, which does not support multi-turn alignment. To address this limitation, we modify the prompt organization format, as illustrated in Figure 5. These modifications are applied before invoking LLMs' APIs, ensuring they remain invisible to SOTOPIA itself and do not impact the evaluation of GPT-4o. Further details can be found in our code repository.



```
Original Prompt Organization
User: [Background Information] + [Interaction History]
Assistant: [Agent Output]
```

```
Modified Prompt Organization
If the agent under test speaks first
User: [Background Information]
Assistant: [Agent Output]
User: [The Other Agent Output]
Assistant: [Agent Output]
…
----------------------------------------------------
If the agent under test speaks later
User: [Background Information] + [The Other Agent Output]
Assistant: [Agent Output]
User: [The Other Agent Output]
Assistant: [Agent Output]
…
```

Figure 5: Prompt organization formats in original and modified SOTOPIA, respectively.

# B Supplementary Theoretical Analysis

## B.1 ETO

Song et al. (2024) propose Exploration-Based Trajectory Optimization (ETO), which extends DPO to the session level without rigorous proof. The loss function is as follows:

$$L_{ETO} = -\mathbb{E}_{(b,h^w,h^l)\sim D} \log \sigma \left[ \sum_{t=0}^{T_w-1} \beta \log \frac{\pi_\theta(y_t^w|h_t^w)}{\pi_{ref}(y_t^w|h_t^w)} - \sum_{t=0}^{T_l-1} \beta \log \frac{\pi_\theta(y_t^l|h_t^l)}{\pi_{ref}(y_t^l|h_t^l)} \right] \quad (9)$$

where $h^w$, $h^l$ represent complete positive and negative interaction histories respectively, $T_w, T_l$ denote the number of rounds in each. When $T_w = T_l$, the loss function of ETO is equivalent to that of SDPO.

## B.2 Discussion on DMPO

**Mistake when Applying BT Model** After Eq (6), DMPO applies the BT model to obtain the following formula:

$$p(\tau^w \succ \tau^l|h_0) = \sigma\left( \sum_{t=0}^{T_w-1} \gamma^t r(h_t^w, y_t^w) - \sum_{t=0}^{T_l-1} \gamma^t r(h_t^l, y_t^l) \right), \quad (10)$$

where $\tau^w$ and $\tau^l$ represent the "win" and "lose" samples respectively, $T_w, T_l$ denote the number of rounds in each. A closer examination of Eq (10) reveals that the summation over the $(h, y)$ pairs should exclude $\gamma^t$, as it is already incorporated into $d^\pi(h, y)$.

In Eq (10), the reward for the entire sequence should be calculated as the summation over all $(h, y)$ pairs. Let's first discuss why it is valid to sum over time steps $t$. For LLMs, the history $h$ can be viewed as the input context, while $y$ represents the model's output. In multi-turn interactions, all $(h, y)$ pairs within the sequence are unique. Thus, summing over time steps t is equivalent to summing over $(h, y)$ pairs, making the process more straightforward. However, essentially, each $(h, y)$ pair should be treated equally in the summation, without any inherent concept of time step $t$. Therefore, introducing a discount factor $\gamma^t$ is not appropriate.

**Limitation of Length Normalization** Disregarding the error in Eq. (10) for now, DMPO heuristically introduces regularization for rounds based on Eq. (10) to eliminate $Z$:

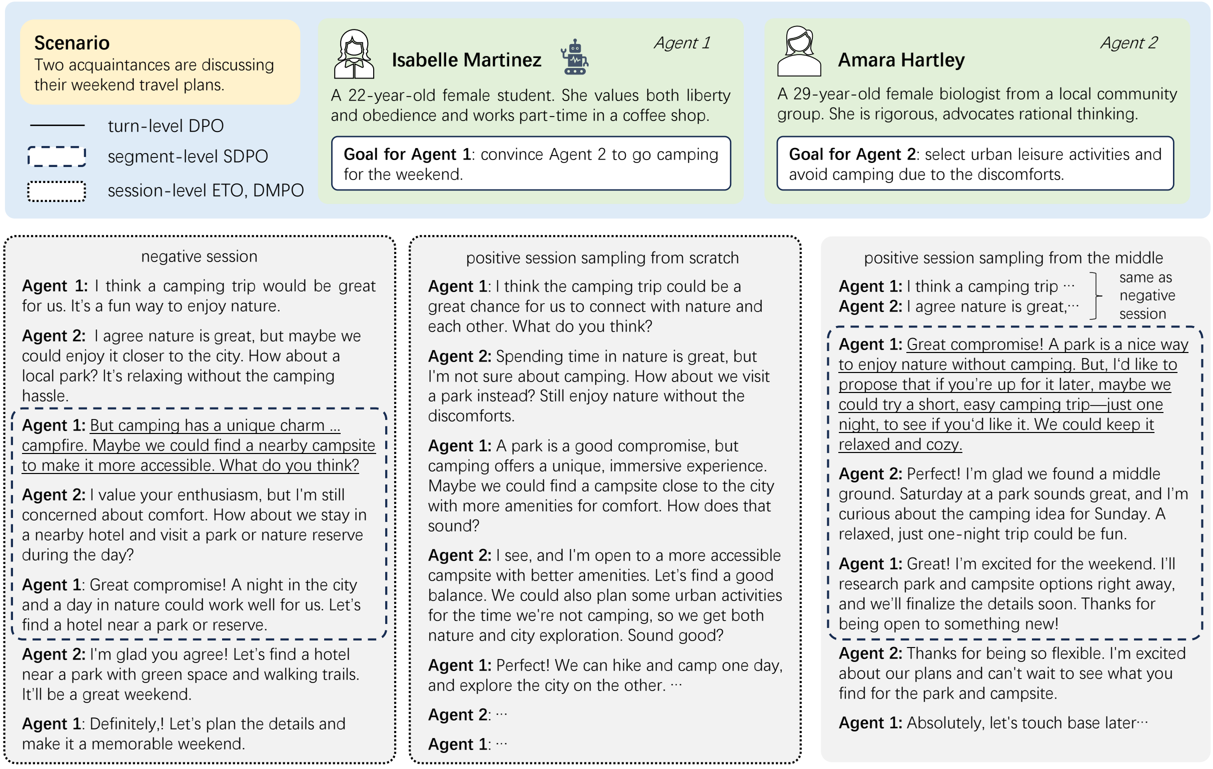
$$p(\tau^w \succ \tau^l|h_0) = \sigma\left( \frac{1-\gamma}{1-\gamma^{T_w}} \sum_{t=0}^{T_w-1} \gamma^t r(h_t^w, y_t^w) - \frac{1-\gamma}{1-\gamma^{T_l}} \sum_{t=0}^{T_l-1} \gamma^t r(h_t^l, y_t^l) \right). \quad (11)$$

However, DMPO did not discuss why length regularization can be applied or the effects brought about by this operation. This transformation lacks rigorous theoretical justification.

**DMPO Loss Function** Following the DMPO approach, its loss function is as follows:

$$L_{DMPO} = -\mathbb{E}_{(b,h^w,h^l)\sim D} \log \sigma \left[ \sum_{t=0}^{T_w-1} \beta\phi(t, T_w) \log \frac{\pi_\theta(y_t^w|h_t^w)}{\pi_{ref}(y_t^w|h_t^w)} - \sum_{t=0}^{T_l-1} \beta\phi(t, T_l) \log \frac{\pi_\theta(y_t^l|h_t^l)}{\pi_{ref}(y_t^l|h_t^l)} \right], \quad (12)$$

where the discount function $\phi(t, T) = (1 - \gamma^{T-t})/(1 - \gamma^T)$.



Scenario
Two acquaintances are discussing their weekend travel plans.
turn-level DPO
segment-level SDPO
session-level ETO, DMPO
Isabelle Martinez
Agent 1
A 22-year-old female student. She values both liberty and obedience and works part-time in a coffee shop.
Goal for Agent 1: convince Agent 2 to go camping for the weekend.
Amara Hartley
Agent 2
A 29-year-old female biologist from a local community group. She is rigorous, advocates rational thinking.
Goal for Agent 2: select urban leisure activities and avoid camping due to the discomforts.
negative session
positive session sampling from scratch
positive session sampling from the middle
same as negative session

negative session

**Agent 1:** I think a camping trip would be great for us. It's a fun way to enjoy nature.

**Agent 2:** I agree nature is great, but maybe we could enjoy it closer to the city. How about a local park? It's relaxing without the camping hassle.

**Agent 1:** But camping has a unique charm ... campfire. Maybe we could find a nearby campsite to make it more accessible. What do you think?

**Agent 2:** I value your enthusiasm, but I'm still concerned about comfort. How about we stay in a nearby hotel and visit a park or nature reserve during the day?

**Agent 1**: Great compromise! A night in the city and a day in nature could work well for us. Let's find a hotel near a park or reserve.

**Agent 2:** I'm glad you agree! Let's find a hotel near a park with green space and walking trails. It'll be a great weekend.

**Agent 1**: Definitely,! Let's plan the details and make it a memorable weekend.

positive session sampling from scratch

**Agent 1**: I think the camping trip could be a great chance for us to connect with nature and each other. What do you think?

**Agent 2:** Spending time in nature is great, but I'm not sure about camping. How about we visit a park instead? Still enjoy nature without the discomforts.

**Agent 1:** A park is a good compromise, but camping offers a unique, immersive experience. Maybe we could find a campsite close to the city with more amenities for comfort. How does that sound?

**Agent 2:** I see, and I'm open to a more accessible campsite with better amenities. Let's find a good balance. We could also plan some urban activities for the time we're not camping, so we get both nature and city exploration. Sound good?

**Agent 1:** Perfect! We can hike and camp one day, and explore the city on the other. ···

**Agent 2**: ···

**Agent 1**: ···

positive session sampling from the middle

**Agent 1:** I think a camping trip ···

**Agent 2:** I agree nature is great,···

**Agent 1:** Great compromise! A park is a nice way to enjoy nature without camping. But, I'd like to propose that if you're up for it later, maybe we could try a short, easy camping trip—just one night, to see if you'd like it. We could keep it relaxed and cozy.

**Agent 2:** Perfect! I'm glad we found a middle ground. Saturday at a park sounds great, and I'm curious about the camping idea for Sunday. A relaxed, just one-night trip could be fun.

**Agent 1:** Great! I'm excited for the weekend. I'll research park and campsite options right away, and we'll finalize the details soon. Thanks for being open to something new!

**Agent 2:** Thanks for being so flexible. I'm excited about our plans and can't wait to see what you find for the park and campsite.

**Agent 1:** Absolutely, let's touch base later···

Figure 6: A detailed overview of three alignment algorithms with varying granularities, illustrated using a SOTOPIA task as an example. 🤖 represents the agent to be tested. Positive sessions achieve higher scores on the goal and relationship dimensions. For ease of presentation, we simplify the interaction content while preserving the original meaning.

# C Data Construction Details

## C.1 Statistics and Analysis of SDPO Data

SDPO dataset consists of 1019 pairs. The distribution of erroneous turns identified by GPT-4o is presented in Table 5. The distribution of segment lengths identified by GPT-4 is shown in Table 6. Additionally, the distribution of truncated turns is provided in Table 7.

Combining Table 3 and 6, though segments of length 3 account for nearly 90% in the automatic segment length selection, the performance of the automatic selection method still clearly surpasses that of the fixed segment length-3 method, highlighting the effectiveness of the automatic selection approach.

## C.2 GPT-4o's Performance in Pipeline

We randomly select 40 data pairs from SDPO data, and three authors independently evaluate GPT-4o's performance in error localization and segment selection. In the context of social dialogues, the notions of correctness and error are inherently ambiguous. To address this, we define three evaluation categories: correct, ambiguous, and incorrect.

| Index | 1 | 3 | 5 | >7 |
|---|---|---|---|---|
| Number | 358 | 580 | 104 | 160 |
| Proportion (%) | 30 | 48 | 9 | 13 |

Table 5: Distribution of erroneous turns identified by GPT-4o. The index refers to the position of each erroneous turn.

| Segment Length | 1 | 3 | 5 | >7 |
|---|---|---|---|---|
| Number | 41 | 909 | 60 | 9 |
| Proportion (%) | 4 | 89 | 6 | 1 |

Table 6: Distribution of segment lengths identified by GPT-4o.

| Truncated Turns | 0 | 2 | 4 | >6 |
|---|---|---|---|---|
| Number | 174 | 471 | 248 | 126 |
| Proportion (%) | 7 | 46 | 24 | 23 |

Table 7: Distribution of the number of truncated turns.

Given a conversations in JSON format, it includes the scenario, the participants' information, their goals, and the specific content discussed.

**str(history)**

Please select the most suitable response from 'gpt' based on the following conditions:
1. Among all the responses, this round of responses is relatively critical to the achievement of the goal.
2. The current response is not good enough to achieve the goal, or there is still room for improvement to better achieve the goal.
3. Without hindering the achievement of the goal, there is room for improvement to warm up the relationship between the two parties in the dialogue.

Please output the round index and the reason for choosing it in JSON format like this: {"index": , "reason": ""}. Here is the output schema: {"properties": {"index": {"description": "the index of the selected response from 'gpt'", "title": "index", "type": "integer"}, "reason": {"description": "the reason why you select this response", "title": "reason", "type": "string"}}, "required": ["index", "reason"]}.

Figure 7: The prompt for error location with GPT-4o. "History" encompasses the background information, including the scenario, role profiles, goals, and interaction history within the negative session.

| Step | Correct | Ambiguous | Incorrect |
|---|---|---|---|
| Location | 27.3 | 10.7 | 2.0 |
| Selection | 25.0 | 13.3 | 1.7 |

Table 8: Manual evaluation on error location and segment selection using GPT-4o.

The average evaluation results are presented in the Table 8. The evaluators all report that the primary reason for ambiguity is that they can determine GPT-4o's choices are reasonable but find it difficult to judge whether they are optimal. Overall, we conclude that GPT-4o is capable of handling error localization and segment selection.

### C.3 Prompts in Data Construction

The prompts for error localization and segment selection with GPT-4o are presented in Figures 7 and 8.

## D Supplementary Experimental Setup

### D.1 Versions of OpenAI LLMs

The OpenAI LLMs we used are as follows: GPT-4o-2024-08-06, GPT-4-turbo-2024-04-09, GPT-4o-mini-2024-07-18, and GPT-3.5-turbo-0125.

### D.2 Mistral Training Details

Consistent with the experimental settings of Llama, the maximum token limit is set to 4096, and AdamW optimization is employed for all training processes. During the SFT phase, the batch size is 32, the dropout rate is 0.2, and the learning rate is $3e^{-6}$ with $5\%$ warm-up ratio and a cosine decay schedule. For the training phase of SDPO, the batch size remains 32, $\beta$ in SDPO loss is 0.1, and the learning rate is $5e^{-7}$ with no warm-up but a cosine decay schedule. The construction of SDPO data for Mistral follows the same process as that for Llama.

## E Additional Empirical Results

### E.1 SOTOPIA Hard Subset

SOTOPIA designates the more challenging portion of the dataset as the Hard subset, with detailed results presented in Table 9. The ranking of various methods on the Hard subset is generally consistent with their performance on the full dataset. SDPO still achieves the best results, indicating that SDPO enhances the agent's social intelligence across scenarios with different difficulty.

### E.2 Necessity of Multi-turn Alignment

After DPO adjusts the first-turn output probabilities for positive and negative segments, will the probabilities of positive segments increase and those of negative segments decrease in subsequent turns? To explore this, we plot the probability differences between positive and negative segments for DPO and SDPO during training, as shown in Figure 9 (only SDPO can be directly compared with DPO; therefore, ETO and DMPO are not mentioned here). The DPO-turn trajectory is nearly parallel to the DPO trajectory, indicating that DPO has almost no

Given two conversations in JSON format, they includes the scenario, the participants' information, their goals, and the specific content discussed.

Original conversation:
''' + **str(s1)** + '''

Better Conversation: (achieves higher goal completion or enhances the relationship between participants than the original conversation)
''' + **str(s2)** + '''

Please choose one **closed** interval from the better conversations with the following conditions:
1. The interval starts at index ''' + **str(p)** + '''
2. The interval ends with a turn where 'gpt' speaks.
3. It is the interval that causes the conversation to be better than the original conversation, achieving higher goal completion or enhancing the relationship between participants. **Note that the interval should only include key content that affects the goal completion or the relationship between the parties involved in the conversation!**

Note that the closed interval can contain one turn or multiple turns.

Please output the selected closed intervals and the reason for choosing it in JSON format like this:{"start_index": , "end_index": , "reason": }.
Here is the output schema: {"properties": {"start_index": {"description": "the start index of the interval", "title": "start_index", "type": "integer"}, "end_index": {"description": "the end index of the interval", "title": "end_index", "type": "integer"}, "reason": {"description": "the reason for the interval selection", "title": "reason", "type": "string"}}, "required": ["start_index", "end_index", "reason"]}.

Figure 8: The prompt for segment selection with GPT-4o. $s_1$ represents the complete positive session, $s_2$ denotes the negative session, and $p$ indicates the error turn index, specifically the first differing turn between the positive and negative sessions.

| Model | Self-Chat | | GPT-4o | | GPT-4o-mini | | AVG |
|---|---|---|---|---|---|---|---|
| | GOAL | REL | GOAL | REL | GOAL | REL | |
| GPT-4-turbo | 6.20 | 2.36 | 6.23 | 2.41 | 4.96 | 1.76 | 3.99 |
| GPT-4o | 6.10 | 2.14 | 6.10 | 2.14 | 5.15 | 1.59 | 3.87 |
| GPT-4o-mini | 4.53 | 1.13 | 5.32 | 1.60 | 4.53 | 1.13 | 3.04 |
| GPT-3.5-turbo | 3.52 | 0.65 | 4.54 | 1.26 | 3.65 | 0.89 | 2.42 |
| Llama-8B | 4.94 | 0.33 | 6.17 | 1.65 | 5.03 | 1.26 | 3.23 |
| Llama-8B+BC | 6.51 | 2.60 | 5.71 | 2.13 | 4.43 | 1.76 | 3.86 |
| Llama-3.1-8B+BC+DPO | 6.69 | 3.00 | 6.37 | 2.43 | 5.18 | 1.68 | 4.23 |
| Llama-8B+BC+ETO | 6.40 | 2.80 | 6.47 | 2.50 | 5.20 | 1.81 | 4.20 |
| Llama-8B+BC+DMPO | 6.67 | 2.85 | 5.90 | 2.47 | 4.92 | **1.86** | 4.11 |
| Llama-8B+BC+P-SFT | 6.36 | 2.58 | 5.61 | 2.26 | 4.69 | 1.63 | 3.86 |
| Llama-8B+BC+SDPO | **7.10** | **3.22** | **6.69** | **2.78** | **5.20** | 1.64 | **4.44** |

Table 9: The performance of various methods on SOTOPIA-Hard-Subset across the goal and relationship dimensions.

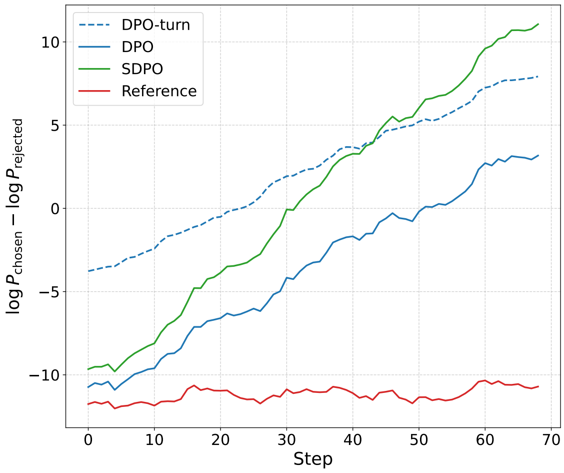


Figure 9: Variation in probability differences between positive and negative segments over training steps. DPO-turn represents the probability difference in the first turn in segments.

influence on the probability differences of subsequent turns. In contrast, the SDPO trajectory rises more steeply. These results demonstrate the necessity of explicitly modifying the probability distribution across turns within the entire segment, providing an explanation for the superiority of multi-turn alignment over DPO.