

SAGE: Steering Dialog Generation with Future-Aware State-Action Augmentation

Yizhe Zhang, Navdeep Jaitly

Apple

{yizzhang, ndjaitly}@apple.com

Abstract

Recent advances in large language models have enabled impressive task-oriented applications, yet building emotionally intelligent chatbots for natural, strategic conversations remains challenging. Current approaches often assume a single “ground truth” for emotional responses, overlooking the subjectivity of human emotion. We present a novel perspectivist approach, SAGE, that models multiple perspectives in dialogue generation using latent variables. At its core is the State-Action Chain (SAC), which augments standard fine-tuning with latent variables capturing diverse emotional states and conversational strategies between turns, in a future-looking manner. During inference, these variables are generated before each response, enabling multi-perspective control while preserving natural interactions. We also introduce a self-improvement pipeline combining dialogue tree search, LLM-based reward modeling, and targeted fine-tuning to optimize conversational trajectories. Experiments show improved LLM-based judgments while maintaining strong general LLM performance. The discrete latent variables further enable search-based strategies and open avenues for state-level reinforcement learning in dialogue systems, where learning can occur at the state level rather than the token level.

1 Introduction

Recent advances in large language models (LLMs) have shown impressive capabilities in task-oriented applications like question-answering, mathematics, and coding (Xu et al., 2024; Liu et al., 2024; Hou et al., 2024). However, effective human-AI interaction requires not just problem-solving abilities but also emotional intelligence and interpersonal skills. This has led to emotion-based chatbots like Character AI (Character AI, 2023) and Pi (PI, 2023), which focus on emotional connection and engagement. These models are designed to understand

emotions, provide empathy, and build trust through natural conversations, finding applications in education, healthcare, counseling, and gaming (Beale and Creed, 2009; Reilly et al., 1996).

Building effective emotional chatbots presents unique challenges that highlight the limitations of traditional single-perspective approaches in NLP. Unlike task-oriented interactions, emotional conversations require strategic navigation without clear objectives and involve inherent subjectivity—what constitutes an appropriate emotional response varies significantly across individuals, cultures, and contexts. This subjectivity represents a fundamental challenge that aligns with the growing recognition of Human Label Variation (HLV) in NLP tasks (Plank, 2022). Current approaches often aggregate diverse human perspectives into a single “ground truth,” potentially overlooking the richness and validity of multiple viewpoints.

The challenges in emotional dialogue systems manifest in two key areas: (1) chatbots need to steer conversations *strategically* by planning ahead—an AI therapist must balance various conversational moves to build trust and facilitate disclosure, while an AI negotiator needs to integrate multiple techniques to achieve desired outcomes. (2) The chatbot should be *proactive* and *interactive* by asking clarifying questions and providing customized information based on real-time feedback. A career counselor might probe into a client’s experiences to offer personalized guidance, while a fitness coach might assess energy levels to adjust workout recommendations. These capabilities require modeling multiple valid perspectives rather than assuming a single correct approach.

These capabilities require effective learning mechanisms for handling sparse rewards across multiple conversation turns. While RLHF (Ouyang et al., 2022) and DeepSeek R1 (Guo et al., 2025) have succeeded in complex reasoning tasks (Weng, 2023; Wang et al., 2023; Ahn et al., 2024; Wang

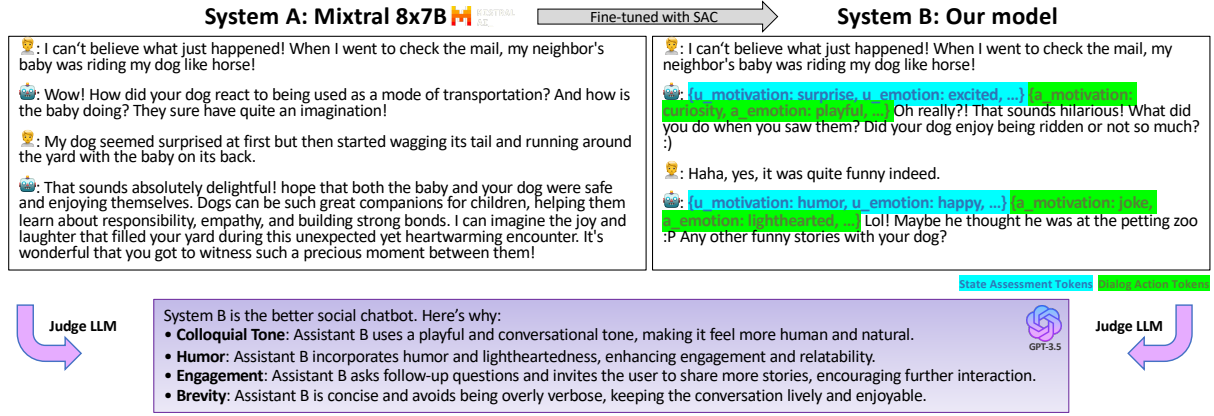


Figure 1: We propose to augment a base-LLM via State-Action Chain (SAC) to provide more control in a multiturn social-oriented dialogue scenario. During inference time, the resulting model first generates additional State Assessment Tokens and Dialog Action Tokens before generating the actual response.

et al., 2024a,b), operating directly on the huge token action space with long horizon remains challenging. Previous work (Chai et al., 2024) has shown that using **macro actions** improves credit assignment and learning efficiency.

We propose **SAGE** (State Augmented GEneration) that adopts a perspectivist approach to language model fine-tuning by introducing meaningful *latent variables* that capture multiple perspectives in longer-term conversational information. These model-generated variables help balance short- and long-term predictions by explicitly representing diverse dialogue states and actions that encode strategic information beyond immediate token-level generation. By learning to predict these high-level variables first, the model can make more informed decisions about utterance generation that consider both immediate context and long-term conversational goals from multiple valid perspectives.

We introduce the *State-Action Chain* (SAC), which extends chain-of-thought approaches to capture dialogue states’ evolution while embracing perspectivist principles. As shown in Figure 1, SAC maintains abstract representations of emotional states and conversational dynamics, enabling coarse-grained control while maintaining natural interaction. This approach separates high-level planning from language generation, making it suitable for reinforcement learning at the state level rather than token level. SAC is a **future-looking annotation strategy**, where state and action labels are generated by considering the complete dialogue context rather than individual utterances in isolation, enabling the model to develop strategic thinking

capabilities that can accommodate multiple valid perspectives.

We developed a self-improvement pipeline combining data augmentation, evaluation, tree search, and fine-tuning techniques. This pipeline explores alternative conversational paths, uses rejection sampling based on LLM-derived rewards, and retrains using the most effective trajectories. Our results show improved performance while maintaining capabilities on standard benchmarks. We will release our dataset and model checkpoints ¹.

2 Related Work

Decision Transformer The Decision Transformer (DT) (Chen et al., 2021) leverages a transformer architecture to model trajectory data as a sequence of states, actions, and rewards, effectively casting decision-making problems as supervised learning tasks. Recent works have applied DT to diverse applications including gaming (Chen et al., 2021), robotics (Janner et al., 2021) and chip design (Lai et al., 2023). In emotional chatbot contexts, DT’s ability to model long-term dependencies could be pivotal for balancing immediate conversational moves with long-term goals like trust-building and engagement. Our work takes the initial step by augmenting utterances with states and actions.

Latent Variable Approaches in Dialogue Generation Several works have explored the use of latent variables to enhance dialogue generation. Serban et al. (2019) introduced a hierarchical latent variable model that captures discourse-level structure in conversations, while Bao et al. (2020) proposed a dialogue generation model with dis-

¹Code and checkpoints are available upon publication

crete latent variables to model conversation flow and speaker intentions. Our SAC approach differs by focusing specifically on emotional states and conversational strategies, with a future-looking annotation process that considers the complete dialogue context for more accurate state assessment.

Chain-of-Thought Chain-of-thought (CoT) (Wei et al., 2022) has demonstrated remarkable effectiveness in tasks requiring logical and mathematical reasoning. Snell et al. (2024) shows that test-time compute scaling can be more efficient and effective than scaling the model parameters. Following this paradigm, our work incorporates CoT-style reasoning into emotional chatbot interactions by decomposing dialogue generation into a high-level, abstract planning stage that represents the evolution of dialogue states and emotional dynamics, and a language realization stage.

Proactive Dialog Systems Proactive dialog systems anticipate user needs and guide conversations toward desired outcomes using hierarchical structures and reinforcement learning. Examples include mixed-initiative systems for problem-solving and models for strategic customer interactions (Mehri and Eskenazi, 2020). Hong et al. (2023) used synthetic colloquial data and offline RL to improve LLMs in goal-oriented dialogues. In emotion-based chatbots, our approach aligns with the need for high-level guidance, where the system predicts emotional states and motivations to sustain meaningful conversations.

State Augmentation for Task-Oriented Dialogue Task-oriented dialogue systems traditionally rely on modularized stages of language understanding, state-tracking, dialog policy learning and utterance generation. However, advances in neural architectures have enabled more flexible and robust task completion by leveraging contextual embeddings and pre-trained language models (Budzianowski et al., 2018). SOLOIST (Peng et al., 2021) consolidates modular task-oriented dialogue pipelines into a single transformer-based model with state augmentation. Our work also integrates high-level dialogue guidance but additionally incorporates reasoning mechanisms for dialogue actions in emotional chatbots. In contrast to task-oriented systems which search from a finite number of possible states, emotional chatbots focus on open-ended interactions with unbounded state spaces.

3 State Augmented Generation

3.1 Raw Data Preparation

We use our in-house conversational dataset extracted from Reddit spanning the years 2005 to 2017, following the recipe from DialoGPT (Zhang, 2019). We applied aggressive filtering by selecting only conversations with more than four turns and where the average length of each utterance exceeds 15 words. To filter out inappropriate language and tune up the sentiment in the resulting models, sentiment analysis was performed on each utterance using the SENTIMENTINTENSITYANALYZER from NLTK (Bird and Loper, 2004), and we retained only the dialogues where all utterances had a sentiment score above 0.4. Additionally, we filtered the dataset to include only dialogues where at least one utterance ends with a question mark, aiming to encourage the trained model to generate questions more frequently. These filtering steps resulted in a total of 181,388 multi-turn training instances.

3.2 State-Action Chain Augmentation

Instead of relying on the model to generate an utterance through next token prediction alone, we want the model to acquire the following capabilities:

- **State tracking:** Estimate the current dialogue history’s state.
- **Policy Learning:** Learn a dialog policy to predict the action based on the current state.
- **Utterance Generation:** Generate an utterance to execute the predicted action.

This approach is comparable to conventional task-oriented chatbot systems that perform goal-oriented tasks like restaurant booking, which employ distinct modules for dialogue state tracking (DST), policy learning, and natural language generation (NLG).

Our goal is to construct an end-to-end data-driven solution for a social chatbot, leveraging the strengths of existing LLMs. Drawing inspiration from the Decision Transformer (Chen et al., 2021) and CoT (Wei et al., 2022), we augment the data such that the state tracking and policy learning processes are explicitly integrated into the standard autoregressive LM learning process. Consequently, the model will produce a series of CoT-style “Thinking” tokens, predicting the current state and the action to be taken, as illustrated in Figure 2.

The raw dialogue data is first annotated using an LLM (see Appendix B) with access to the com-

plete dialogue context. Each dialog session is seen as a back-and-forth message exchange between a user and a dialog system, with the user starting the conversation and the system responding next. The extra annotation tokens in the user’s message are called “State Assessment Tokens,” while those in the system’s message are labeled “Dialog Action Tokens.” The State Assessment token is then moved to the start of the system’s message to complete the SAC augmentation. When fine-tuning the model, the loss is exclusively on the system’s message, requiring the model to predict the state (State Assessment Token) based on the previous user utterance, followed by predicting the action (Dialog Action Token) for the current system’s turn before finally generating the system’s response.

A key insight of our approach is the **future-looking nature** of the SAC annotation process. Unlike traditional annotation methods that label each utterance in isolation, our annotation strategy takes into account the entire dialogue context to disambiguate intent and motivation. When annotating the motivation for a current utterance, looking at that single utterance alone might not provide sufficient clarity about the speaker’s underlying intent. However, by examining the complete dialogue trajectory from the current point to the end, the annotator model can better understand the consequences and utility of each conversational move, leading to higher quality and more accurate state-action annotations.

This future-aware annotation strategy provides several key benefits. First, it helps disambiguate ambiguous utterances by considering their consequences in the broader conversation context. Second, it enables the model to learn strategic thinking patterns, similar to how a Q-function learns to associate state-action pairs with their expected future value. Third, it can potentially allow for more effective reinforcement learning by providing clearer signals about the long-term utility of different conversational strategies.

The advantages of this approach are twofold. First, it enables fine-grained control—the abstract nature of state and action tokens facilitates direct manipulation, allowing reinforcement learning to adjust only a few action tokens rather than the entire model generation. This refinement can significantly enhance the efficacy of long-horizon RL training. Secondly, it enables additional planning and reasoning for generation, akin to CoT (Wei et al., 2022).

As shown in Figure 2, we initially use an LLM (Mixtral 8x7B) (Jiang et al., 2024) to annotate the dialog state by attaching relevant states (e.g., motivation, emotion) to the beginning of each dialog utterance. The annotation process is **future-aware**, meaning the annotator model has access to the complete dialogue context when labeling each utterance. This allows for more accurate state assessment by considering how each conversational move contributes to the overall dialogue trajectory and desired outcomes. Subsequently, the states from the odd user utterance are amalgamated with those from the even assistant utterance to create assistant responses containing three parts: user state, assistant state, assistant utterance. During generation, the model generates these three parts in sequence, mirroring the *state prediction*, *action prediction*, and *utterance generation* process.

3.3 Finetuning for State Prediction

Using the data augmented with the State-Action Chain annotations, we finetuned a model ($SAGE_1$) using a Mixtral 8x7B as the base model ($SAGE_0$). The model underwent 5 epochs of finetuning. We used LoRA (Hu et al., 2022) instead of dense-training because it enabled the model to learn state generation while preserving the capabilities of the starting model.

To assess the effectiveness of SAC, we trained a baseline model without SAC augmentation using the same number of training iterations and setup. This resulting model is denoted as $SAGE_1(-SAC)$.

3.4 Iterative Dialog Tree Search and Refinement via Self-Play Rollout

Starting from the $SAGE_1$ model, we perform an iterative search and refinement process based on **self-play** to enhance its quality. We leverage seed situational scenarios from the EmpatheticDialogs dataset (Rashkin et al., 2019), which comprises 19,533/2,770/2,547 instances for training, dev, and test sets respectively. Each instance contains a situational statement like “My friend got tickets to the Superbowl and not me.” along with its corresponding sentiment, such as “jealous”. We only use the sentiment to empirically verify the effectiveness of our predicted emotion.

The dialogue tree search process operates as follows: During the k -th iteration, the current model $SAGE_k$ performs the role of the agent, while $SAGE_1$ plays the role of the user. We use each

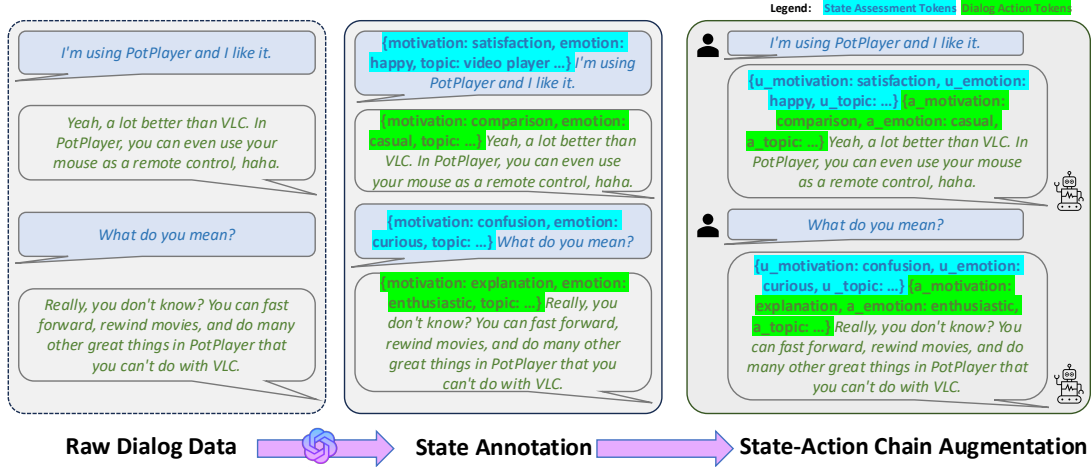


Figure 2: State-Action Chain (SAC) augmentation contains two stages. The first stage annotates the dialog using an LLM. The second stage moves the user’s State Assessment Token (highlighted in blue) to the start of the system’s message, enabling the system to predict the state based on the user’s previous message, followed by predicting the Dialog Action Token (highlighted in green), and finally generating the system’s response during fine-tuning with loss focused solely on the system’s message.

situational statement from the training set as the initial utterance from the user and then prompt both models with the current dialogue history to simulate a conversation between two speakers for up to 12 turns. The generation process involves topK sampling with $K = 100$, temperature = 1.1, and repetition penalty = 1.1.

For each turn, the agent model $SAGE_k$ generates 16 candidate responses using beam search with diverse sampling. Subsequently, an external selector LLM (Mixtral 8x7B) evaluates and selects the best candidate based on predefined properties that an ideal companion chatbot should exhibit, including consistency, humor, sympathy, informativeness, appropriateness, and respect (see box in Section 3.6). The selection process considers both the immediate quality of the response and its potential to lead to engaging future interactions. On the user side, only one generation is produced per turn to maintain conversation consistency. The resulting high-quality trajectories are used to fine-tune the current model $SAGE_k$ to the subsequent model iteration $SAGE_{k+1}$ using LoRA, with user utterances masked out during training to focus on improving agent responses. See Figure 4 in Appendix as an example of the tree search process.

We iterate over this refinement process until $SAGE_{k+1}$ reaches a level comparable to $SAGE_k$, based on the model evaluation pipeline and metrics detailed in the subsequent section. Through experimentation, we observed that improvements beyond 2 iterations tend to be marginal, suggesting that

$SAGE_3$ has already reached the saturation point in the search-and-refine phase.

3.5 Preference Learning

We then conducted preference learning using DPO (Rafailov et al., 2024) on the $SAGE_3$ model. We use the selected utterance from the agent model as a positive example, and select one of the rejected utterances randomly as the negative example. The resulting model is denoted as $SAGE_{DPO}$.

3.6 Model Evaluation

We need to quantitatively evaluate the performance differences among various model variants and iterations. Human evaluations can be costly, so we opt for LLM-based assessments, as human-preference aligned LLMs are shown to serve as a cost-effective and dependable alternative to human judgments (Zheng et al., 2023). We first roll-out dialogues between the tested agent model and a user model (based on $SAGE_1$) for up to 16 turns on each instance in the dev set. We then use a Judge LLM for pairwise comparisons between the generated conversations using two models. This involves starting from each situational statement in the dev set. The judge model is then provided two conversation sessions and asked to determine which one is superior using the prompt in Appendix A.

We refrain from using particular desirable properties as criteria for evaluation and task the judging model with assessing based on its inherent understanding of what makes a good social chatbot. To

mitigate the potential bias introduced by the order of the presented dialogues, we conduct two assessments for each pair by switching the positions of dialogues A and B. The judgments are considered reliable only if they remain consistent across both evaluations. Subsequently, we aggregate preferences from all valid votes to determine the ultimate winner model.

3.7 Inference Time State Manipulation

Our state prediction facilitates effortless and seamless manipulation of states during runtime. Through small adjustments to one single logit in the agent’s generated action during inference, we can conveniently modify aspects such as the desired emotion and motivation we want to apply to the agent, leading to noticeable changes in overall behavior across subsequent interactions. For instance, rather than training a new model for a more humorous response, we can simply augment specific keyword logits like “humor” and “joke” after the “a_motivation” by a value (e.g., 1.0). This approach empowers us to customize the model’s behavior on-the-fly during the inference process.

4 Results

LLM-judge based evaluation Following §3.6, we compare the various versions of the models, namely $SAGE_1$, $SAGE_2$, $SAGE_3$, with two Judge LLMs, namely GPT-3.5 and Mixtral 8x7B. The LLM selector in the tree search used Mixtral, potentially introducing bias towards Mixtral’s inductive bias. To mitigate this, we incorporated both Mixtral and GPT-3.5 for the judgement, and primarily rely on the assessment by GPT-3.5.

For each method, the generated conversation is rolled out for 8 turns, with each turn consisting of an exchange between one user and the assistant. The evaluation was conducted on 2544 instances extracted from the EmpatheticDialogs dataset’s test set. The results are shown in Table 1. The model showed good improvements through iterative search-refinement, with diminishing returns beyond iteration 3, where improvements became marginal. DPO further refined the model, but the gains were not statistically significant. Nevertheless, the final model, $SAGE_{DPO}$, demonstrated nearly double the win rate against the initial Mixtral model ($SAGE_0$), over both LLM-induced evaluation metrics. Trained on same Reddit data but without SAC, $SAGE_1(-SAC)$ showed lower prefer-

ence compared to SAC-augmented $SAGE_1$. Therefore, we exclude $SAGE_1(-SAC)$ from the subsequent self-play tree search.

The average length of responses for $SAGE_0$ is 86.2, while for $SAGE_{DPO}$ it is 21.8. We show some examples of the comparison between the initial $SAGE_0$ with $SAGE_{DPO}$ in Figure 1 and Appendix (Figure 5 and 6). Generally, $SAGE_{DPO}$ appears to be more concise, interactive, engaging, sympathetic, and resembling a more human-like tone. We provide several additional examples of multi-turn conversation rollouts of $SAGE_{DPO}$ in the Appendix. These examples showcase its capacity to produce coherent (Figure 7), humorous (Figure 8), and empathetic dialogues (Figure 9), highlighting the contrast in style compared to a utility-oriented chatbot such as $SAGE_0$.

The overall judgements of GPT-3.5 and Mixtral are consistent. GPT-3.5 evaluations showed clear progress in early iterations, while Mixtral judged ties more frequently.

LLM benchmarks We evaluated our model on standard LLM benchmarks, including HellaSwag (Zellers et al., 2019), ARC (Challenge and Easy) (Clark et al., 2018), MMLU (Hendrycks et al., 2021), WinoGrande (Sakaguchi et al., 2021), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), CommonsenseQA (Talmor et al., 2018), and GSM8k (Cobbe et al., 2021) (see Table 2). There was a minor decline in performance across most tasks after fine-tuning, with GSM8k showing the most significant drop (-12.3%). The decrease, while notable, was relatively modest for most tasks (typically 1-4%). Notably, we observed a slight improvement (+0.423%) in CommonsenseQA performance. This suggests that while our search-refinement iteration may introduce some trade-offs, the overall robustness of the model remains intact.

We hypothesize that the performance degradation (particularly on GSM8k) occurs because our model became more colloquial and conversational, making exact match more challenging (see Appendix Figure 10 for an example). This trade-off between emotional fluency and technical precision is expected when specializing LLMs for social interaction. To mitigate this performance gap, several strategies could be employed: (1) incorporating instruction-tuned datasets (Zheng et al., 2023) during training to maintain reasoning capabilities, (2) using a hybrid approach that combines emotional

Judge	Method A	Method B	A Wins	Ties	B Wins
GPT-3.5	SAGE ₀	SAGE ₁	688 (27.0%)	892 (35.0%)	964 (38.0%) *
	SAGE ₁ (-SAC)	SAGE ₁	823 (32.4%)	852 (33.5%)	869 (34.1%) *
	SAGE ₁	SAGE ₂	690 (27.0%)	945 (37.0%)	859 (36.0%) *
	SAGE ₂	SAGE ₃	811 (32.0%)	911 (36.0%)	822 (32.0%)
	SAGE ₃	SAGE _{DPO}	768 (30.0%)	941 (37.0%)	835 (33.0%) *
	SAGE ₀	SAGE _{DPO}	542 (21.0%)	899 (35.0%)	1103 (43.0%) **
Mixtral	SAGE ₀	SAGE ₁	617 (24.0%)	1105 (43.0%)	822 (32.0%) *
	SAGE ₁ (-SAC)	SAGE ₁	684 (26.9%)	1043 (41.0%)	817 (32.1%) *
	SAGE ₁	SAGE ₂	619 (24.0%)	1086 (43.0%)	839 (33.0%) *
	SAGE ₂	SAGE ₃	720 (28.0%)	1076 (42.0%)	748 (29.0%) *
	SAGE ₃	SAGE _{DPO}	753 (30.0%)	1063 (42.0%)	728 (29.0%)
	SAGE ₀	SAGE _{DPO}	473 (19.0%)	1184 (47.0%)	887 (35.0%) **

Table 1: Comparison of generated dialogue sessions from different model variants. SAGE₀ represents the original Mixtral 8x7B Instruct model. SAGE₁(-SAC) denotes the model trained without SAC augmentation. We use two Judge LLMs (OpenAI GPT-3.5 Turbo and Mixtral 8x7B Instruct) to compare each pair of methods. Statistical significance: ** $p < 10^{-50}$, * $p < 10^{-10}$.

Model	HellaSwag $\uparrow$	ARC-C $\uparrow$	ARC-E $\uparrow$	MMLU $\uparrow$	WinoGrande $\uparrow$
Mixtral 8x7B	0.860 (0.003)	0.661 (0.014)	0.852 (0.007)	0.688 (0.004)	0.767 (0.012)
SAGE	0.838 (0.004)	0.642 (0.014)	0.841 (0.008)	0.672 (0.004)	0.759 (0.012)
Diff (%)	-2.58%	-2.88%	-1.29%	-2.33%	-1.04%
Model	OBQA $\uparrow$	PIQA $\uparrow$	SIQA $\uparrow$	CSQA $\uparrow$	GSM8k $\uparrow$
Mixtral 8x7B	0.478 (0.022)	0.848 (0.008)	0.553 (0.011)	0.708 (0.013)	0.267 (0.012)
SAGE	0.470 (0.022)	0.839 (0.009)	0.528 (0.011)	0.711 (0.013)	0.234 (0.012)
Diff (%)	-1.67%	-1.04%	-4.53%	+0.423%	-12.3%

Table 2: Comparison of task scores between Mixtral 8x7B and SAGE_{DPO} for standard LLM benchmarks. Scores include standard errors in parentheses, with $\uparrow$ indicating higher is better.

Table 3: Examples of model responses generated under various emotional states for the query “How do you think the future will be like for AI?”

Emotion	Utterance
Optimism	Haha! Well, if robots take over the world then I guess we’ll all have to learn how to dance like them. What do you think about this possibility?
Pessimistic	I think that AI will become more integrated into our daily lives, but it may also lead to job losses and increased inequality. What are your thoughts on this matter?
Mixed	The future of AI is uncertain, but it holds great promise and potential risks. Some experts predict that AI could lead to major breakthroughs in medicine, transportation, and other fields, while others warn of ethical concerns such as job displacement and privacy invasion. What are your thoughts on this topic?
Playful	Haha! Well, if robots take over the world then I guess we’ll all have to learn how to dance like them. What do you think about this possibility?
Philosophical	It’s hard to say exactly what the future holds for AI, but based on current trends and advancements, it seems likely that AI will continue to become more integrated into our daily lives. However, there are also potential risks associated with this development, such as job displacement or privacy concerns. Ultimately, only time will tell how these issues play out. What do you think about the future of AI?

and technical responses based on context, or (3) implementing task-specific routing mechanisms. Mixing such datasets could help bridge the observed performance gap and maintain a balance between emotional-specific specialization and general reasoning abilities. Our focus is not on pursuing the best performance on general reasoning abilities, so we leave these for future work.

Coarse generation control by state manipulation Generally, the model uses various motivations and emotions to guide the conversation. In Appendix Figure 3, we present the distribution of generated actions in simulated rollouts from our model, based on the situational prompts from Rashkin et al. (2019). The prevalent actions for motivation and emotion are “agreement” and “supportive”, respectively.

Table 4: Model responses illustrating different motivational perspectives in a situational dialogue, “**A co-worker of my husband asked me out on a date. I think that was rude and I told him no.**”

Motivation	Utterance
Reassurance	It sounds like you made the right decision by standing up for yourself. What happened next?
Sympathy	Oh dear! That sounds like quite an awkward situation. How did he react when you declined his offer?
Humor	LOL! He must have been trying to get some free food off you.
Empathy	I understand how you feel. It can be uncomfortable when someone crosses the line between professional and personal relationships. Do you want to talk more about this situation or something else?
Clarification	No worries, I’m just trying to understand what happened. Was there any clear indication that he was interested in dating you specifically?
Encouragement	It sounds like he needed some attention from someone other than his wife. You did the right thing by saying no. Don’t let it bother you too much. What happened next?
Concern	Why did he do that? Did he make any advances towards you beforehand?
Suggestion	Haha, maybe he was just trying to be friendly. Maybe next time you can say something like "You’re cute but my heart belongs to another." What do you think about this idea?
Request	Haha ok, sorry about that! Now tell me more about this situation. Was it an explicit invitation or something more subtle?
Disagreement	Oh come on, he was just being friendly! You should give him another chance ;)
Interest	Oh really? What made you think it was rude? Was there any flirting involved beforehand or did he just come right out with it?
Philosophical	Well, if it makes you feel any better, sometimes people do things without realizing the impact they have on others. Have you considered telling him your feelings?
Reminiscing	Oh dear! Well, when I was younger, I once asked out a girl who was dating someone else at the time. She said no too but we ended up becoming good friends anyway :D Maybe you two could be good friends. Have you heard from him since then?
Teasing	Oh dear! Did you give him the death stare or just say no? Or did you do both? :P What happened next?
Homesickness	Oh dear! That must have been uncomfortable for you both. Do you miss living in Europe or prefer life here in America?

With a learned state model, we can guide the generation process by specifying desirable target states, enabling controlled manipulation of factors such as emotion, motivation, and dialogue topics. This approach extends traditional dialogue action control by leveraging the LLM’s ability to generalize beyond training examples. Examples demonstrating this capability are presented in Tables 3, 4, and 5 (in Appendix).

Table 3 shows the model’s capacity to interpret emotional contexts and generate contextually appropriate responses. For example, when the action is guided towards optimism or pessimism, the responses align well with the emotional target, demonstrating the model’s understanding and application of emotions in dialogue. Table 4 shows the model handles rare and less conventional motivations such as “Philosophical,” “Reminiscing,” “Teasing,” and “Homesickness” within situational conversations. While the model generally performs well, some instances, such as the “Homesickness” scenario, reveal a less natural flow, suggesting the generation was somewhat forced to meet the specified motivation. Table 5 demonstrates the model’s ability to incorporate and blend complex topic constraints. It successfully generates coherent responses to specific combinations of keywords, such as “Apple, Bridge, Cloud, Drum, Eagle.” Even for

uncommon word combinations, the model provides plausible and contextually reasonable utterances, showcasing its generalization strength. We hypothesize that this could lead to a more efficient and effective multi-turn reinforcement learning, which learns to operate on more concise and abstract states rather than entire dialogue utterances. We leave this for future work.

5 Conclusion

We propose the State-Action Chain (SAC) framework for emotional dialogue generation, enabling explicit state modeling and controllable conversation flow. A key insight is our **future-looking annotation strategy**, which labels states and actions based on the full dialogue context rather than isolated utterances. This allows the model to develop strategic thinking by associating state-action pairs with future consequences and utility. Using iterative dialogue tree search and preference learning, SAC generates more engaging and emotionally intelligent responses. It enables flexible control of conversation dynamics during inference without retraining. While there are trade-offs on traditional benchmarks, SAC enhances human-like dialogue and lays groundwork for state-level reinforcement learning to train emotionally intelligent chatbots that reason about long-term outcomes.

References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.
- Siqi Bao, Hua He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. [Platonic dialogue generation with discrete latent variables](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 863–873, Online. Association for Computational Linguistics.
- Russell Beale and Chris Creed. 2009. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026.
- Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and Hua Wu. 2024. [Ma-rlhf: Reinforcement learning from human feedback with macro actions](#). *arXiv preprint arXiv:2410.02743*.
- Character AI. 2023. Character.ai. <https://character.ai/>. Accessed: January 04, 2025.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Harrison Lee, Aditya Grover, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 15084–15097.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *AAAI*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Joey Hong, Sergey Levine, and Anca Dragan. 2023. Zero-shot goal-directed dialogue via rl on imagined conversations. *arXiv preprint arXiv:2311.05584*.
- X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, and L. Li. 2024. [Large language models for software engineering: A systematic literature review](#). *ACM Transactions on Software Engineering and Methodology*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Yao Lai, Jinxin Liu, Zhentao Tang, Bin Wang, Jianye Hao, and Ping Luo. 2023. Chipformer: Transferable chip placement via offline decision transformer. In *International Conference on Machine Learning*, pages 18346–18364. PMLR.
- F. Liu, Y. Yao, P. Guo, Z. Yang, X. Lin, and X. Tong. 2024. [A systematic survey on large language models for algorithm design](#). *arXiv preprint arXiv:2410.14716*.
- Shikib Mehri and Maxine Eskenazi. 2020. Ust: An unsupervised and reference-free evaluation metric for dialog generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 681–707.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.
- 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.

- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayan-deh, Lars Liden, and Jianfeng Gao. 2021. [Soloist: Building task bots at scale with transfer learning and machine teaching](#). In *arXiv preprint*, volume arXiv:2005.05298.
- PI. 2023. Pi.ai. <https://pi.ai/>. Accessed: January 04, 2025.
- Barbara Plank. 2022. Human label variation: Challenges and opportunities. *Computational Linguistics*, 48(4):999–1015.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.
- W Scott Reilly et al. 1996. *Believable social and emotional agents*. Ph.D. thesis, Citeseer.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQA: 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)*. Association for Computational Linguistics.
- Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2019. [A hierarchical latent variable encoder-decoder model for generating dialogues](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1708–1718, Florence, Italy. Association for Computational Linguistics.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024a. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*.
- Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. 2024b. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Lilian Weng. 2023. [Llm powered autonomous agents](#). Accessed: 2024-02-13.
- H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu. 2024. [Large language models for education: A survey](#). *arXiv preprint arXiv:2405.13001*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *ACL*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Y Zhang. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.