Deep Reinforcement Learning with Hierarchical Action Exploration for Dialogue Generation

Itsugun Cho[1], **Ryota Takahashi**[1], **Yusaku Yanase**[1], **Hiroaki Saito**[1]

Keio University, Japan[1]

Keio University

1858

CALAMVS GLADIO FORTIOR

Why use Reinforcement Learning

Being equipped with a "forward-looking" capacity in the dialogue to enhance user experience.

Problem Scenario

Policy π : **Agent**
State s : **Context**
Action a : **Response** $a \sim \pi(a|s)$
Reward r : **Reward** $r = R(s, a)$
State s' : **Next Context** $s' = T(s, a)$

State-value function : $V^{\pi}(s)$
Action-value function : $Q^{\pi}(s, a)$

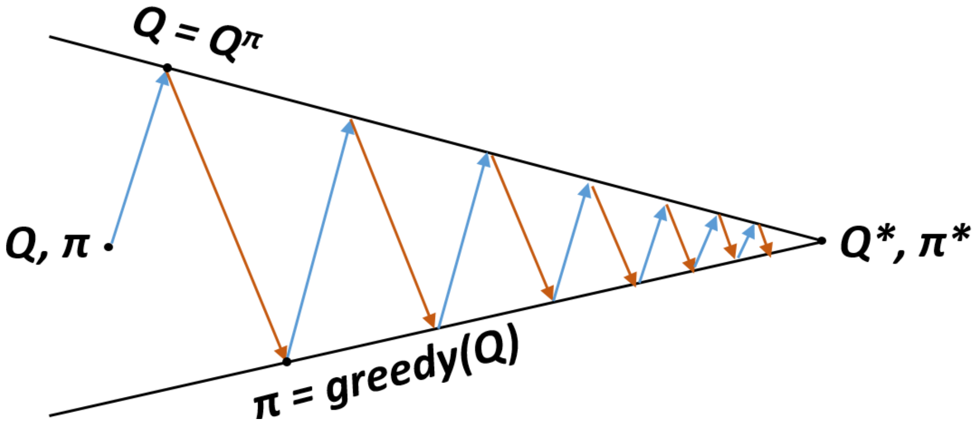
Policy Iteration (Actor-critic)

Approximate Dynamic Programming

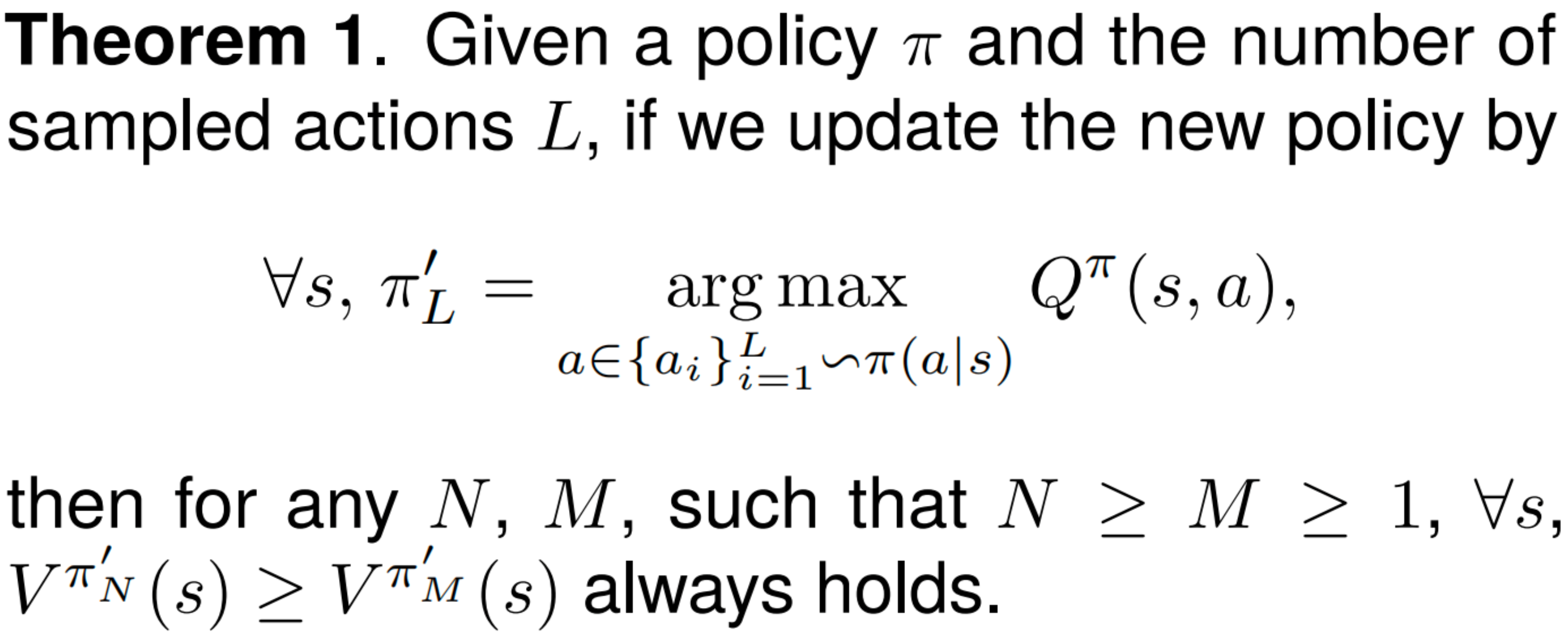
$$\text{Evaluation}: Q \leftarrow \arg\min_{Q} \mathbb{E}_{(s,a,r,s')\backsim\mathcal{D}} \left[(r + \gamma \mathbb{E}_{a'\backsim\pi(a'|s')}[Q(s',a')] - Q(s,a))^2\right]$$

$$\text{Improvement}: \pi \leftarrow \arg\max_{\pi} \mathbb{E}_{s\backsim\mathcal{D},a\backsim\pi(a|s)} \left[Q(s,a)\right]$$

Policy Improvement Theorem



Theorem 1

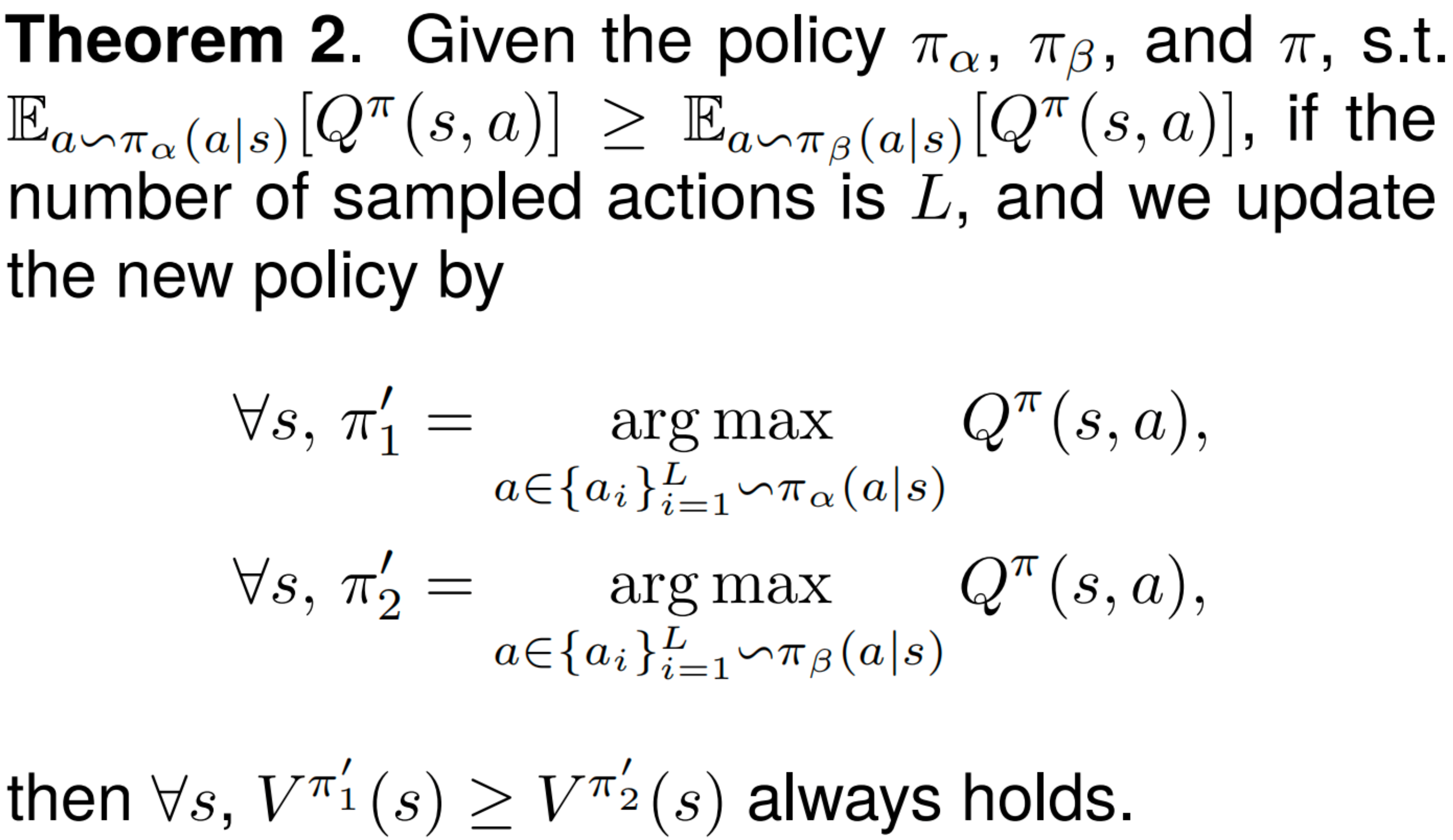


Theorem 1. Given a policy π and the number of sampled actions L, if we update the new policy by

$$\forall s,\ \pi'_L = \underset{a \in \{a_i\}_{i=1}^{L} \backsim \pi(a|s)}{\arg\max} Q^{\pi}(s, a),$$

then for any N, M, such that $N \geq M \geq 1$, $\forall s$, $V^{\pi'_N}(s) \geq V^{\pi'_M}(s)$ always holds.

Theorem 2



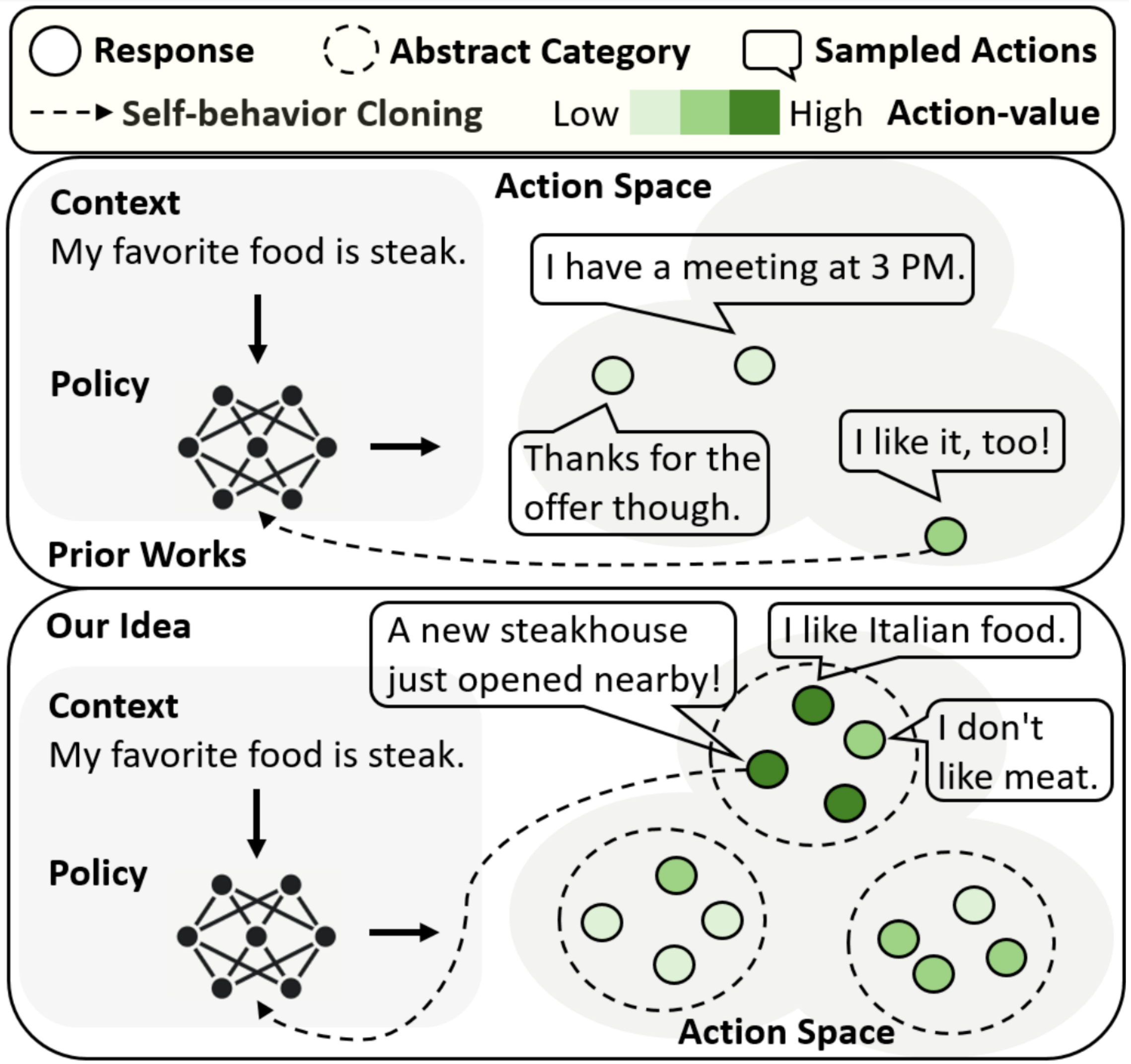
Theorem 2. Given the policy π_α, π_β, and π, s.t. $\mathbb{E}_{a \backsim \pi_\alpha(a|s)}[Q^\pi(s,a)] \geq \mathbb{E}_{a \backsim \pi_\beta(a|s)}[Q^\pi(s,a)]$, if the number of sampled actions is L, and we update the new policy by

$$\forall s,\ \pi'_1 = \underset{a \in \{a_i\}_{i=1}^{L} \backsim \pi_\alpha(a|s)}{\arg\max} Q^\pi(s,a),$$

$$\forall s,\ \pi'_2 = \underset{a \in \{a_i\}_{i=1}^{L} \backsim \pi_\beta(a|s)}{\arg\max} Q^\pi(s,a),$$

then $\forall s$, $V^{\pi'_1}(s) \geq V^{\pi'_2}(s)$ always holds.

Dual-granularity Q-function



Rewards

The average cosine similarity between the agent’s response and dull responses.

The outpouring of the emotion of surprise.

The length of response.

Asking questions.

Corpus and Agent

DailyDialog	Dialogues
Training set	11,118
Testing set	1,000
Validation set	1,000

Agent	Descriptions
GPT-2	proposed by Radford et al. (2019), is an unsupervised autoregressive language model for textual generation.
DialoGPT	is a pre-trained dialogue model proposed by Zhang et al. (2020). This model is based on GPT-2, using the Reddit comments dataset.
T5	is a unified framework proposed by Raffel et al. (2020) that converts all text based language tasks into a text-to-text format via the transfer learning technique.
GODEL	is a pre trained dialogue model proposed by Peng et al. (2022). This model is based on T5, using the Reddit discussion dataset.

Automatic Metrics

CS: The average **c**osine **s**imilarity between the agent's response and dull responses

SE: The outpouring of the **e**motion of **s**urprise

LR: The **l**ength of **r**esponse

AQ: **A**sking **q**uestions

Human Metrics

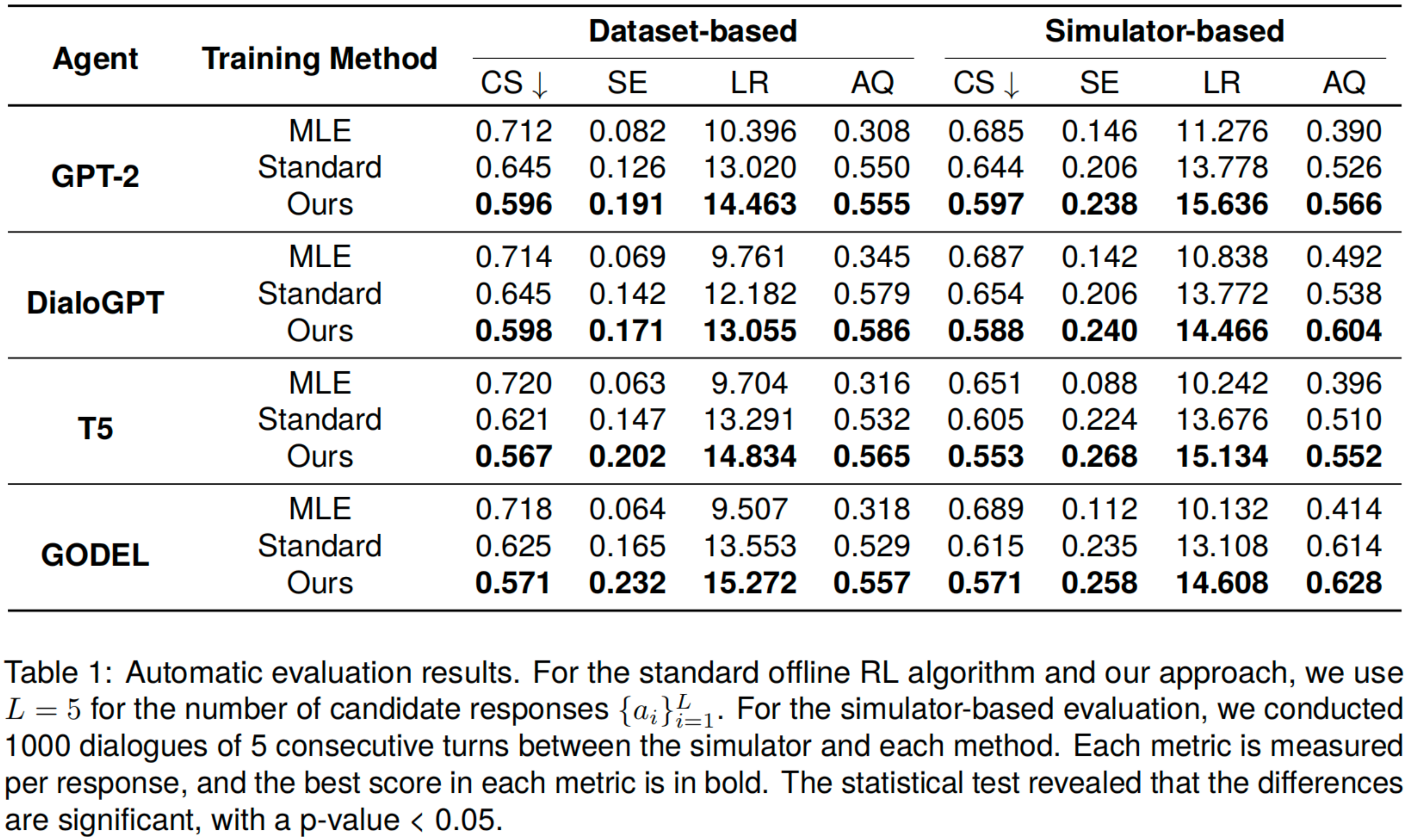
Quality measures the coherence and grammatical accuracy of the agents’ responses

Informativeness measures the diversity and hallucination of the agents’ responses

Empathy measures the degree to which agents respond with concern or affectivity

Engagingness measures the desire to engage the agents in a long conversation

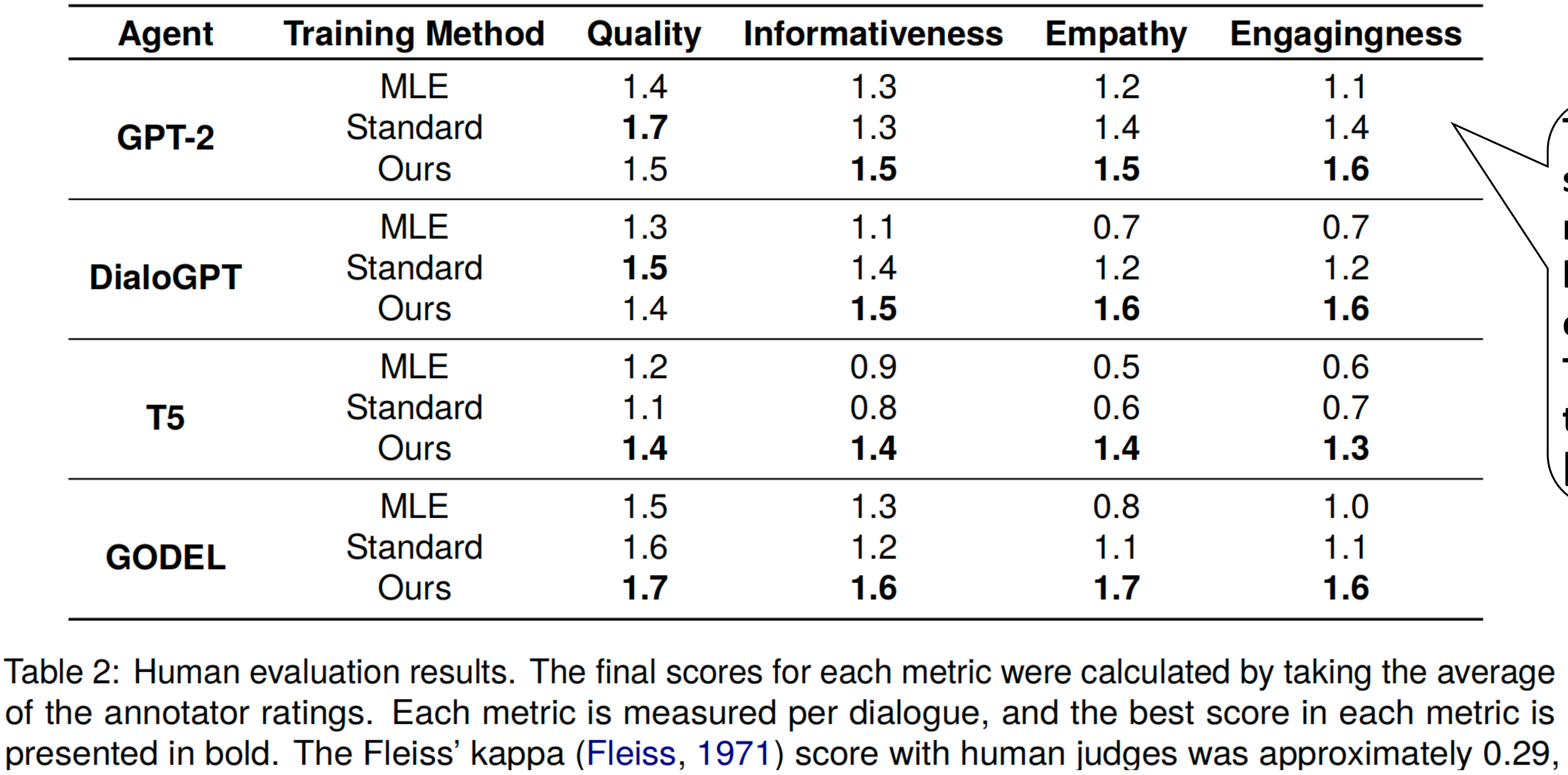
Experiments



Agent	Training Method	Dataset-based				Simulator-based			
		CS ↓	SE	LR	AQ	CS ↓	SE	LR	AQ
GPT-2	MLE	0.712	0.082	10.396	0.308	0.685	0.146	11.276	0.390
	Standard	0.645	0.126	13.020	0.550	0.644	0.206	13.778	0.526
	Ours	**0.596**	**0.191**	**14.463**	**0.555**	**0.597**	**0.238**	**15.636**	**0.566**
DialoGPT	MLE	0.714	0.069	9.761	0.345	0.687	0.142	10.838	0.492
	Standard	0.645	0.142	12.182	0.579	0.654	0.206	13.772	0.538
	Ours	**0.598**	**0.171**	**13.055**	**0.586**	**0.588**	**0.240**	**14.466**	**0.604**
T5	MLE	0.720	0.063	9.704	0.316	0.651	0.088	10.242	0.396
	Standard	0.621	0.147	13.291	0.532	0.605	0.224	13.676	0.510
	Ours	**0.567**	**0.202**	**14.834**	**0.565**	**0.553**	**0.268**	**15.134**	**0.552**
GODEL	MLE	0.718	0.064	9.507	0.318	0.689	0.112	10.132	0.414
	Standard	0.625	0.165	13.553	0.529	0.615	0.235	13.108	0.614
	Ours	**0.571**	**0.232**	**15.272**	**0.557**	**0.571**	**0.258**	**14.608**	**0.628**

Table 1: Automatic evaluation results. For the standard offline RL algorithm and our approach, we use $L = 5$ for the number of candidate responses $\{a_i\}_{i=1}^{L}$. For the simulator-based evaluation, we conducted 1000 dialogues of 5 consecutive turns between the simulator and each method. Each metric is measured per response, and the best score in each metric is in bold. The statistical test revealed that the differences are significant, with a p-value $<$ 0.05.

Experiments

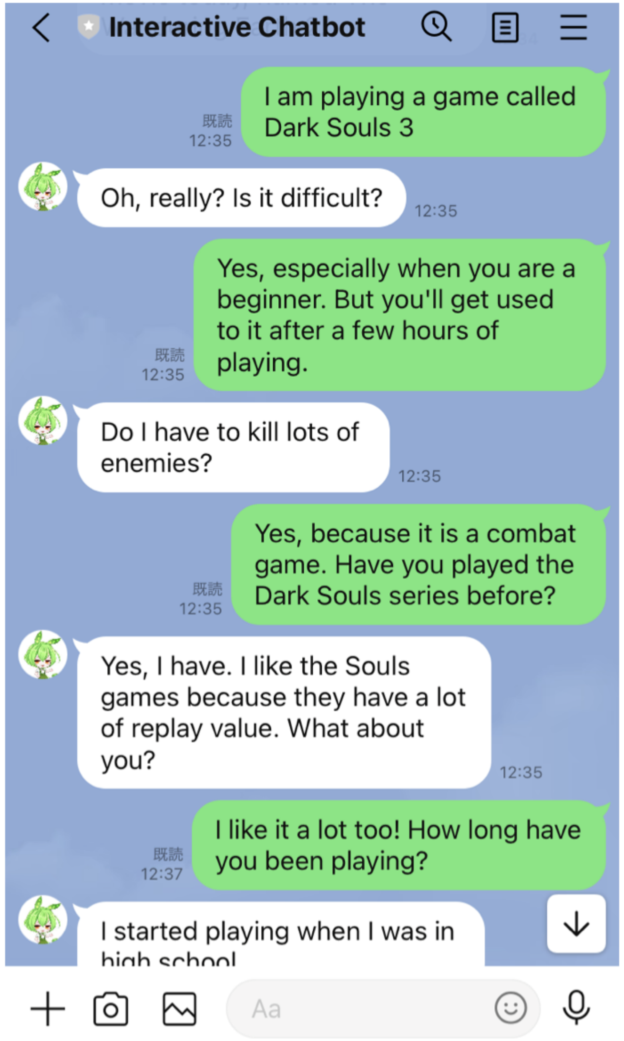


Agent	Training Method	Quality	Informativeness	Empathy	Engagingness
GPT-2	MLE	1.4	1.3	1.2	1.1
	Standard	**1.7**	1.3	1.4	1.4
	Ours	1.5	**1.5**	**1.5**	**1.6**
DialoGPT	MLE	1.3	1.1	0.7	0.7
	Standard	**1.5**	1.4	1.2	1.2
	Ours	1.4	**1.5**	**1.6**	**1.6**
T5	MLE	1.2	0.9	0.5	0.6
	Standard	1.1	0.8	0.6	0.7
	Ours	**1.4**	**1.4**	**1.4**	**1.3**
GODEL	MLE	1.5	1.3	0.8	1.0
	Standard	1.6	1.2	1.1	1.1
	Ours	**1.7**	**1.6**	**1.7**	**1.6**

Table 2: Human evaluation results. The final scores for each metric were calculated by taking the average of the annotator ratings. Each metric is measured per dialogue, and the best score in each metric is presented in bold. The Fleiss' kappa (Fleiss, 1971) score with human judges was approximately 0.29,

Ten native speakers were recruited for human evaluation. The scale of these metrics is [0, 1, 2].

Details regarding Interactive



I am playing a game called Dark Souls 3
既読 12:35

Oh, really? Is it difficult? 12:35

Yes, especially when you are a beginner. But you'll get used to it after a few hours of playing.
既読 12:35

Do I have to kill lots of enemies? 12:35

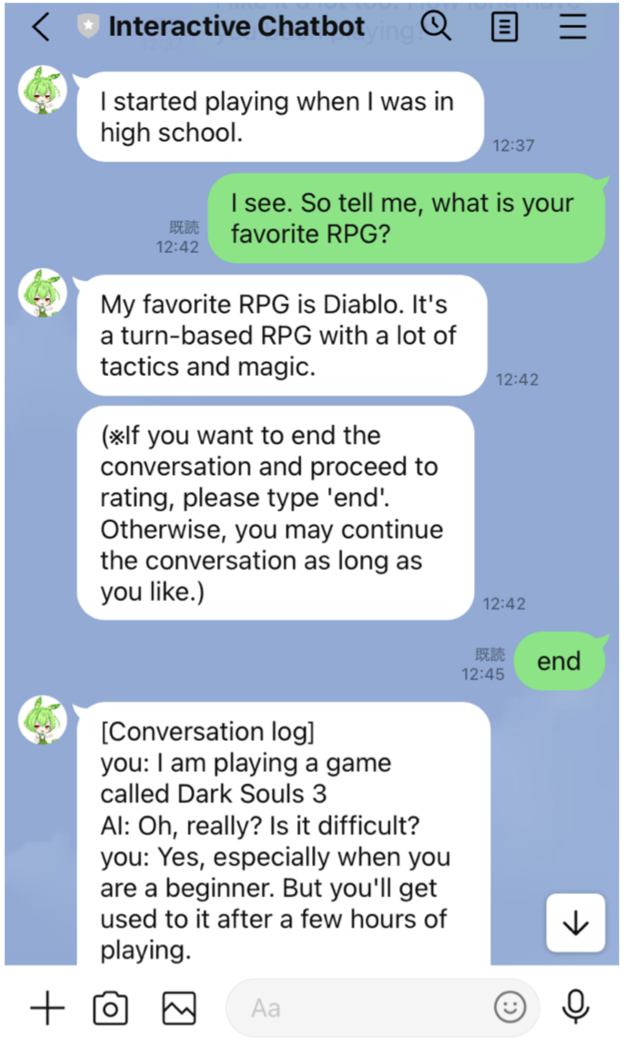
Yes, because it is a combat game. Have you played the Dark Souls series before?
既読 12:35

Yes, I have. I like the Souls games because they have a lot of replay value. What about you? 12:35

I like it a lot too! How long have you been playing?
既読 12:37

I started playing when I was in high school

(a) Chat Interface 1



I started playing when I was in high school. 12:37

I see. So tell me, what is your favorite RPG?
既読 12:42

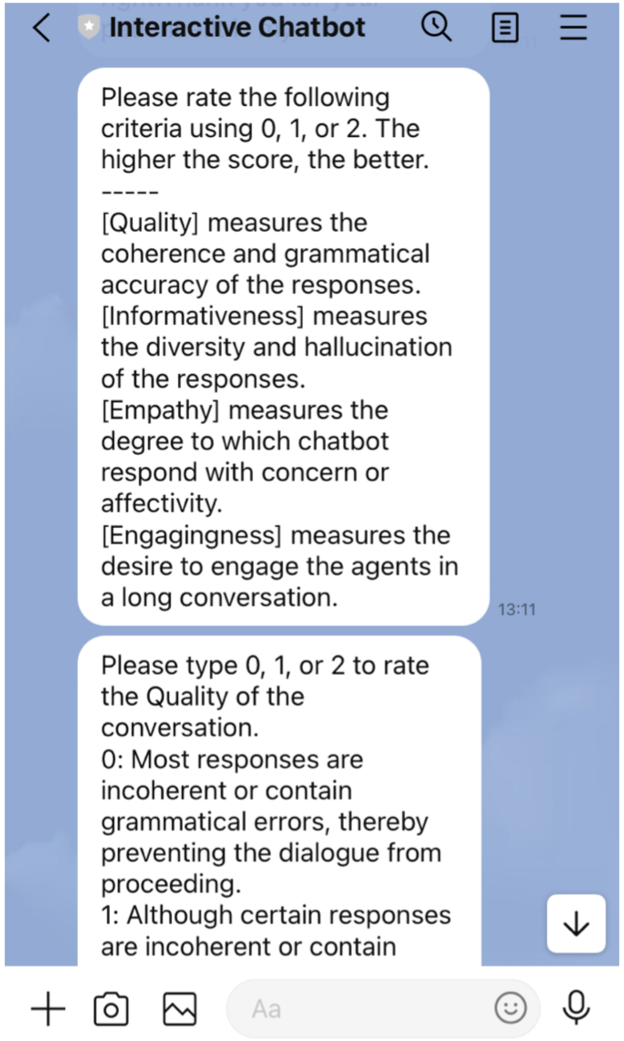
My favorite RPG is Diablo. It's a turn-based RPG with a lot of tactics and magic. 12:42

(※If you want to end the conversation and proceed to rating, please type 'end'. Otherwise, you may continue the conversation as long as you like.) 12:42

end
既読 12:45

[Conversation log]
you: I am playing a game called Dark Souls 3
AI: Oh, really? Is it difficult?
you: Yes, especially when you are a beginner. But you'll get used to it after a few hours of playing.

(b) Chat Interface 2



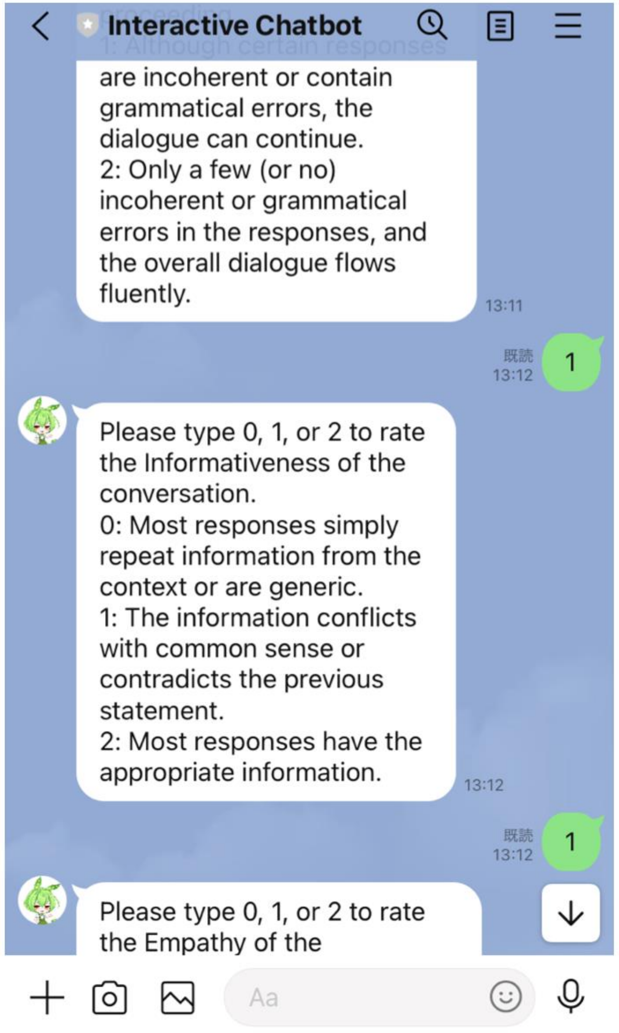
Please rate the following criteria using 0, 1, or 2. The higher the score, the better.

[Quality] measures the coherence and grammatical accuracy of the responses.
[Informativeness] measures the diversity and hallucination of the responses.
[Empathy] measures the degree to which chatbot respond with concern or affectivity.
[Engagingness] measures the desire to engage the agents in a long conversation. 13:11

Please type 0, 1, or 2 to rate the Quality of the conversation.
0: Most responses are incoherent or contain grammatical errors, thereby preventing the dialogue from proceeding.
1: Although certain responses are incoherent or contain

(c) Ratings Interface 1

Details regarding Interactive



proceeding
1: Although certain responses are incoherent or contain grammatical errors, the dialogue can continue.
2: Only a few (or no) incoherent or grammatical errors in the responses, and the overall dialogue flows fluently.
13:11

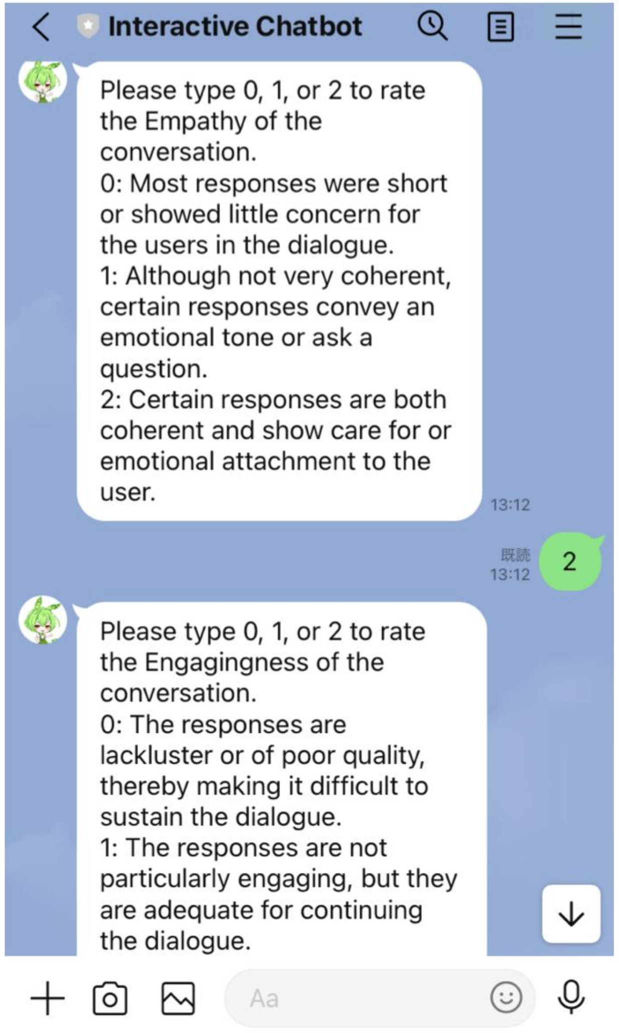
既読 13:12 1

Please type 0, 1, or 2 to rate the Informativeness of the conversation.
0: Most responses simply repeat information from the context or are generic.
1: The information conflicts with common sense or contradicts the previous statement.
2: Most responses have the appropriate information.
13:12

既読 13:12 1

Please type 0, 1, or 2 to rate the Empathy of the

(d) Ratings Interface 2

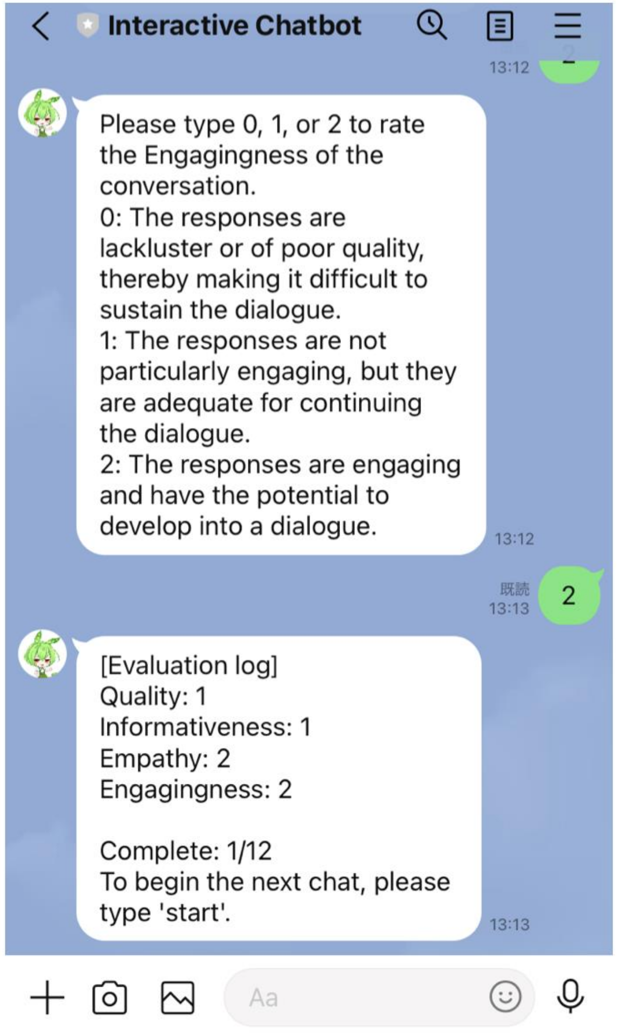


Please type 0, 1, or 2 to rate the Empathy of the conversation.
0: Most responses were short or showed little concern for the users in the dialogue.
1: Although not very coherent, certain responses convey an emotional tone or ask a question.
2: Certain responses are both coherent and show care for or emotional attachment to the user.
13:12

既読 13:12 2

Please type 0, 1, or 2 to rate the Engagingness of the conversation.
0: The responses are lackluster or of poor quality, thereby making it difficult to sustain the dialogue.
1: The responses are not particularly engaging, but they are adequate for continuing the dialogue.

(e) Ratings Interface 3



13:12

Please type 0, 1, or 2 to rate the Engagingness of the conversation.
0: The responses are lackluster or of poor quality, thereby making it difficult to sustain the dialogue.
1: The responses are not particularly engaging, but they are adequate for continuing the dialogue.
2: The responses are engaging and have the potential to develop into a dialogue.
13:12

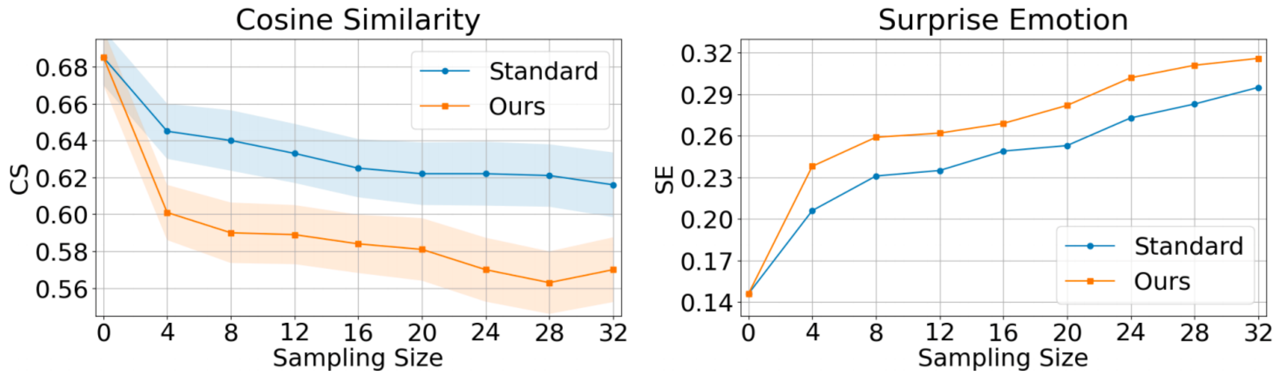
既読 13:13 2

[Evaluation log]
Quality: 1
Informativeness: 1
Empathy: 2
Engagingness: 2

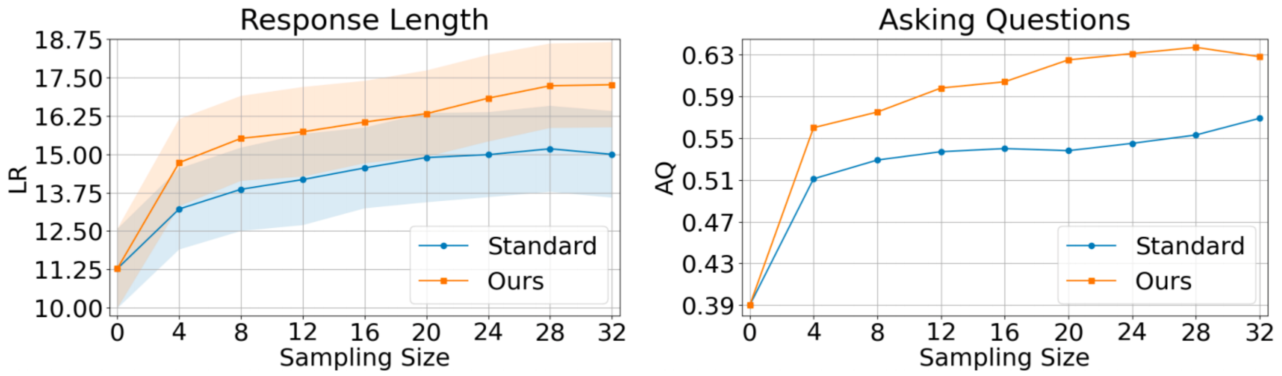
Complete: 1/12
To begin the next chat, please type 'start'.
13:13

(f) Ratings Interface 4

Further Verification



(a) CS (b) SE



(c) LR (d) AQ

Figure 3: The evolution of the agent's performance for each metric with the increased sampling size. The scale for the X-axis is a multiple of 4, and 0 represents the MLE without RL. Bands indicate half a standard deviation.

Further Verification

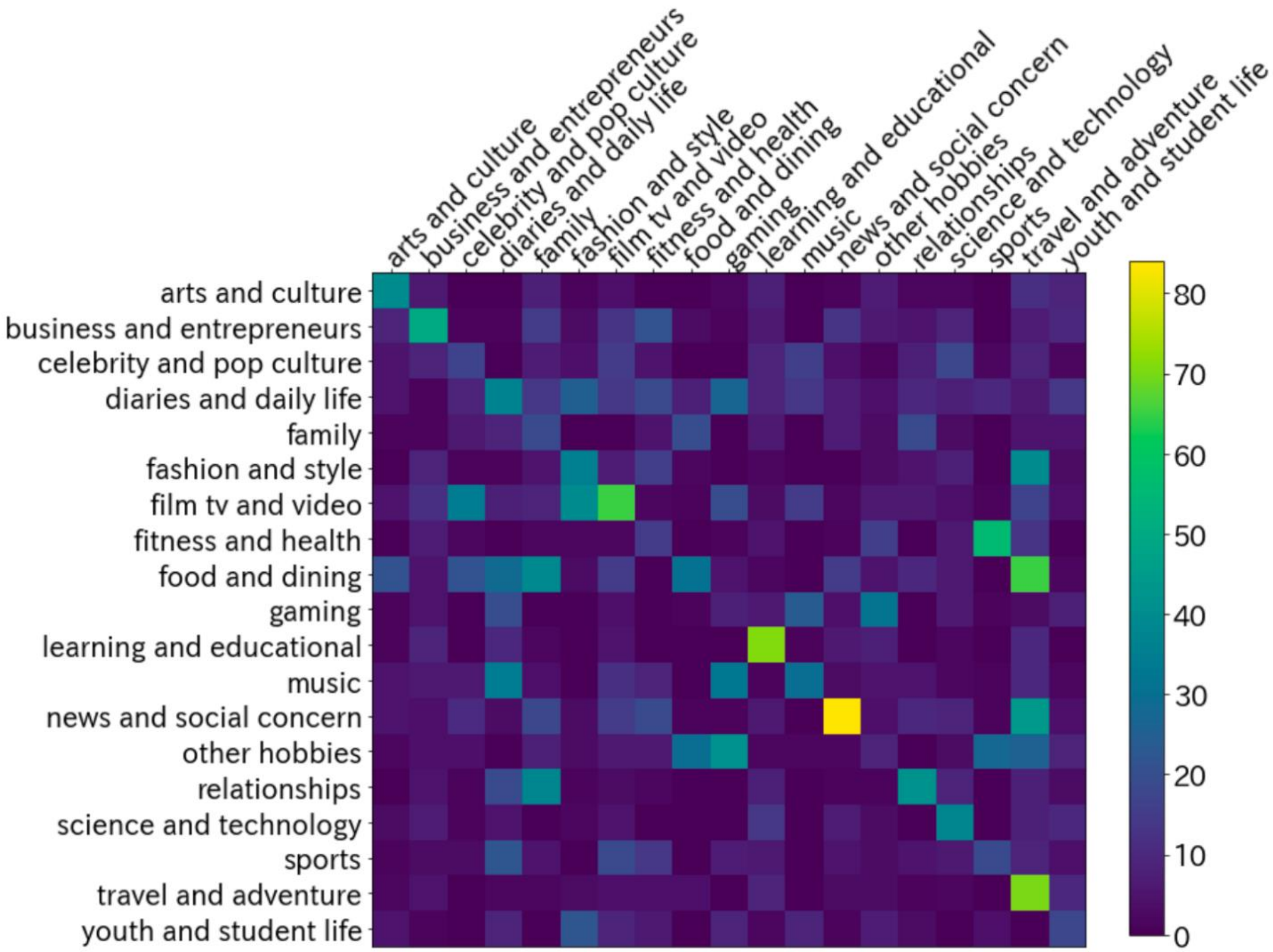
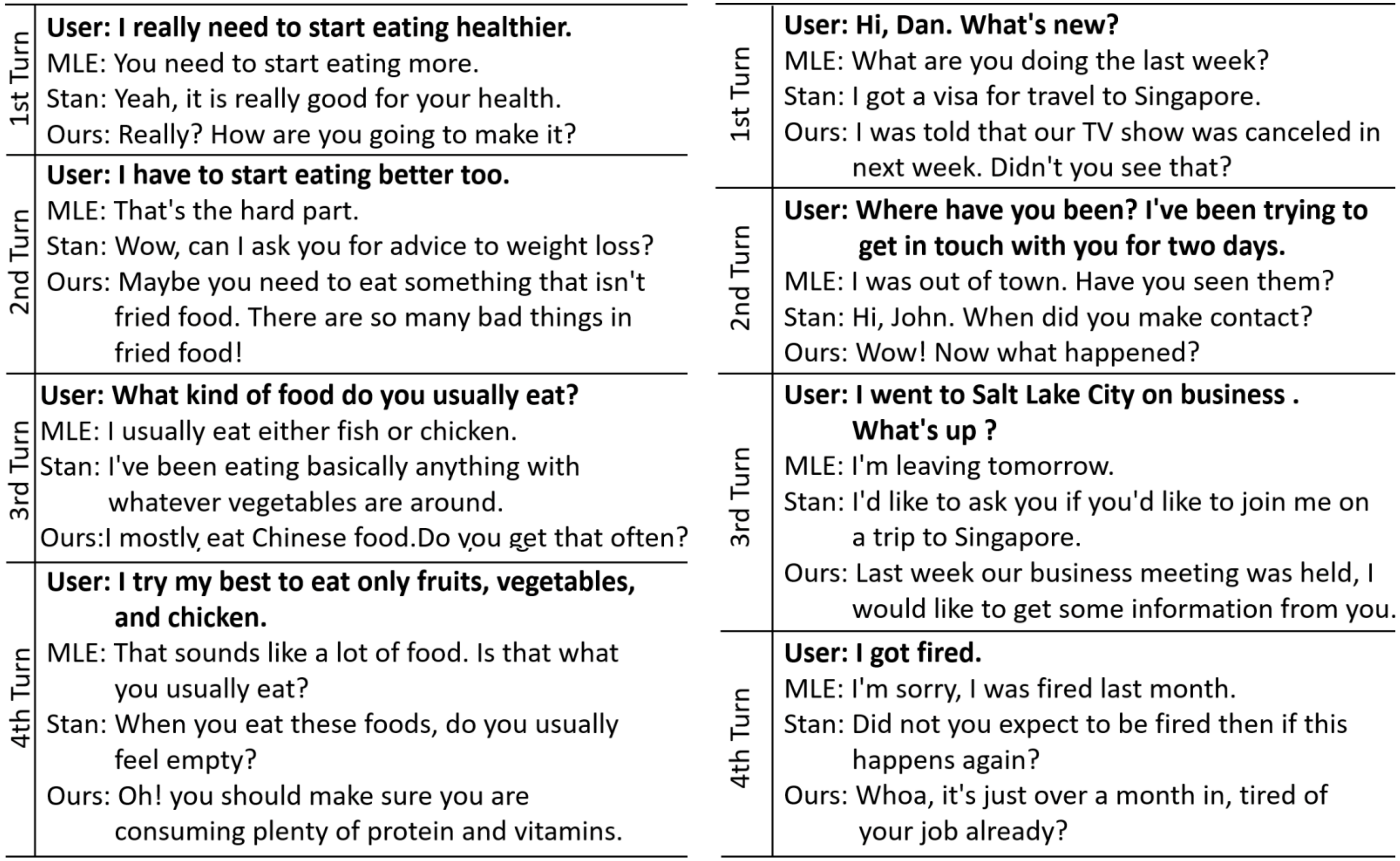


Figure 4: Behavioral analysis of coarse-grained Q-function. The label on the Y-axis represents the topic of each context, whereas the label on the X-axis represents the selected action category. The color bar represents the number of occurrences of the corresponding selection.

Further Verification



1st Turn	**User: I really need to start eating healthier.** MLE: You need to start eating more. Stan: Yeah, it is really good for your health. Ours: Really? How are you going to make it?
2nd Turn	**User: I have to start eating better too.** MLE: That's the hard part. Stan: Wow, can I ask you for advice to weight loss? Ours: Maybe you need to eat something that isn't fried food. There are so many bad things in fried food!
3rd Turn	**User: What kind of food do you usually eat?** MLE: I usually eat either fish or chicken. Stan: I've been eating basically anything with whatever vegetables are around. Ours:I mostly eat Chinese food.Do you get that often?
4th Turn	**User: I try my best to eat only fruits, vegetables, and chicken.** MLE: That sounds like a lot of food. Is that what you usually eat? Stan: When you eat these foods, do you usually feel empty? Ours: Oh! you should make sure you are consuming plenty of protein and vitamins.

(a) DialoGPT

1st Turn	**User: Hi, Dan. What's new?** MLE: What are you doing the last week? Stan: I got a visa for travel to Singapore. Ours: I was told that our TV show was canceled in next week. Didn't you see that?
2nd Turn	**User: Where have you been? I've been trying to get in touch with you for two days.** MLE: I was out of town. Have you seen them? Stan: Hi, John. When did you make contact? Ours: Wow! Now what happened?
3rd Turn	**User: I went to Salt Lake City on business . What's up ?** MLE: I'm leaving tomorrow. Stan: I'd like to ask you if you'd like to join me on a trip to Singapore. Ours: Last week our business meeting was held, I would like to get some information from you.
4th Turn	**User: I got fired.** MLE: I'm sorry, I was fired last month. Stan: Did not you expect to be fired then if this happens again? Ours: Whoa, it's just over a month in, tired of your job already?

(b) T5

Figure 2: Case study. DialoGPT and T5 are used as the agents in the case study. For brevity, the standard method is abbreviated as stan. The ground truth for responding to each user utterance is the next user utterance in the dialogue.

Acknowledgement

This work was supported by JST SPRING, Grant Number JPMJSP2123

Thank you for your listening!