SIGIR 2025
Padova
ITALY

Out-of-Boundary Query Mitigation
Universal Knowledge Boundary

Yang Deng

Singapore Management University

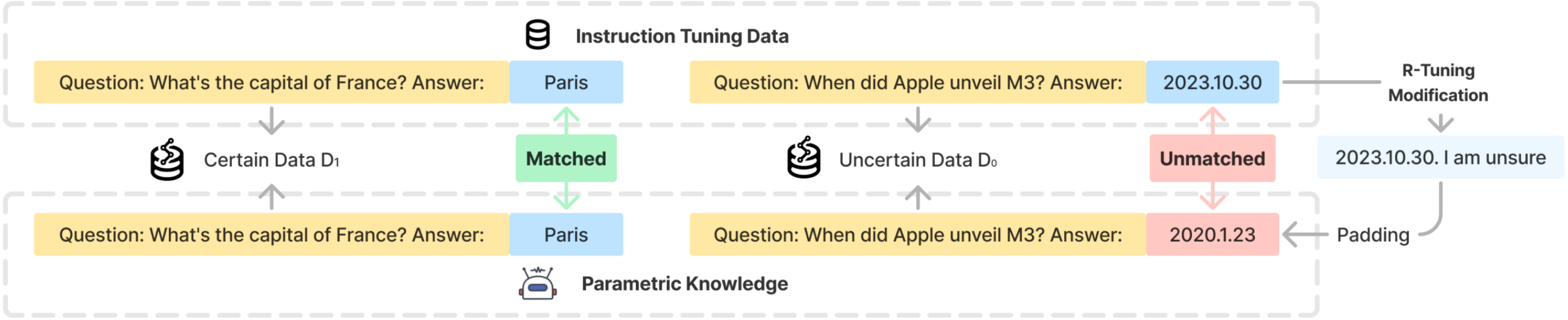
Mitigation of Model-Agnostic Unknown Knowledge

- **Refusal or Abstention**
  - Refusal Fine-tuning
  - Uncertainty-based Reinforcement Learning
  - Self-alignment
- **Ask Clarification Questions**
  - In-Context Learning
  - Reinforcement Learning
  - Preference Optimization

Mitigation of Model-Agnostic Unknown Knowledge

- ❑ **Refusal or Abstention**
  - ❑ Refusal Fine-tuning
  - ❑ Uncertainty-based Reinforcement Learning
  - ❑ Self-alignment
- ❑ **Ask Clarification Questions**
  - ❑ In-Context Learning
  - ❑ Reinforcement Learning
  - ❑ Preference Optimization

Refusal-Aware Instruction Tuning (R-Tuning)



- ❑ **Refusal-Aware Data Identification**

The question with mismatch between the prediction and the ground-truth label results

- ❑ **Refusal-Aware Data Construction**

Construct template-based refusal responses, e.g., "I am unsure"

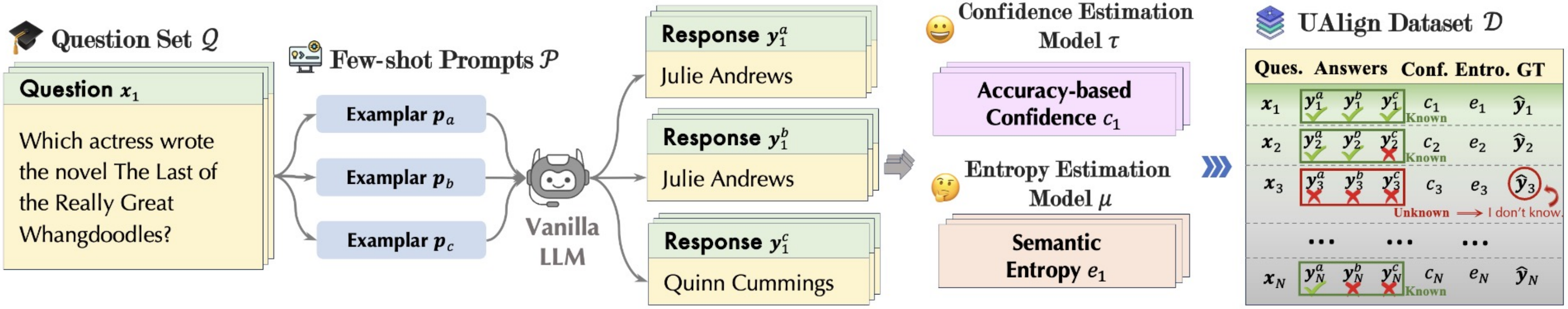
- ❑ **Supervised Fine-tuning**

Zhang et al., "R-Tuning: Instructing Large Language Models to Say 'I Don't Know'?" (NAACL '24)

Mitigation of Model-Agnostic Unknown Knowledge

- ❏ **Refusal or Abstention**
  - ❏ Refusal Fine-tuning
  - ❏ Uncertainty-based Reinforcement Learning
  - ❏ Self-alignment
- ❏ **Ask Clarification Questions**
  - ❏ In-Context Learning
  - ❏ Reinforcement Learning
  - ❏ Preference Optimization

Uncertainty-based Alignment (UAlign)

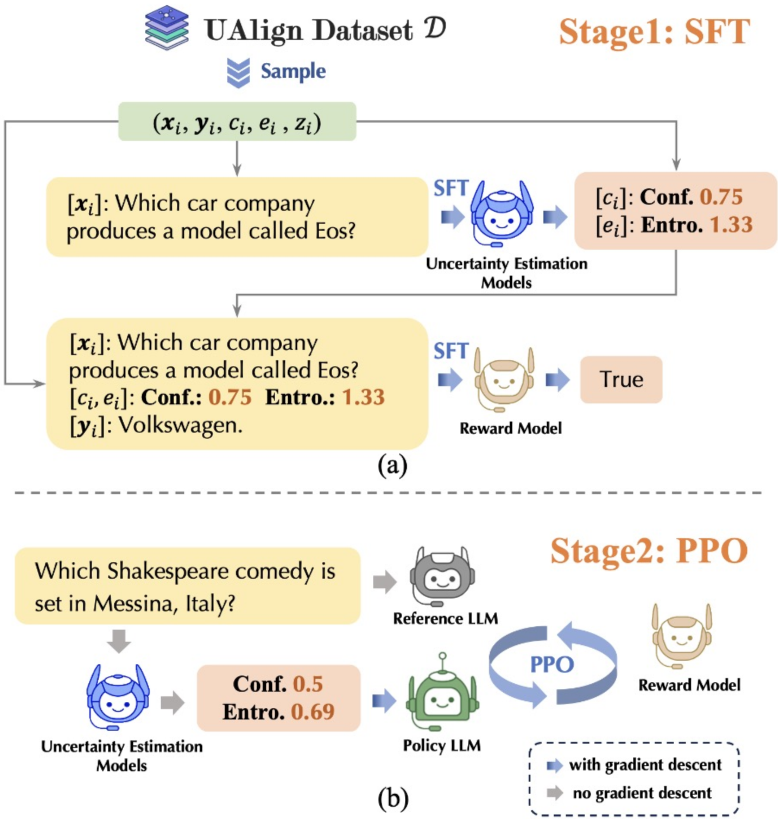


UAlign Data Construction

- ❑ Response Sampling
- ❑ Uncertainty Measurement: Accuracy-based Confidence & Semantic Entropy

Xue et al., "UALIGN: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models" (ACL '25)

Uncertainty-based Alignment (UAlign)



UAlign Training Framework

- **Supervised Fine-tuning** to train uncertainty estimation model
- **Reward Model Training** to train a reward model as a binary evaluator to determine if a generated answer is correctly conditioned on the question, confidence, and entropy.
- **PPO Alignment** to optimize the LLM's factual expressions to a question with the uncertainty measurements.

Xue et al., "UALIGN: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models" (ACL '25)

Mitigation of Model-Agnostic Unknown Knowledge

- ❑ **Refusal or Abstention**
  - ❑ Refusal Fine-tuning
  - ❑ Uncertainty-based Reinforcement Learning
  - ❑ Self-alignment
- ❑ **Ask Clarification Questions**
  - ❑ In-Context Learning
  - ❑ Reinforcement Learning
  - ❑ Preference Optimization

Issues of Refusal

Q: What animal can be found at the top of the men's Wimbledon trophy?

A: The answer is unknown. — **Unknown Question Detection**

A: The question is incorrect. — **Unknown Question Classification**

Not User-friendly;
Fail to Meet User Information Needs

How to properly respond to unknown questions?

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Issues of Refusal

Q: What animal can be found at the top of the men's Wimbledon trophy?

A: The answer is unknown. — **Unknown Question Detection**

A: The question is incorrect. — **Unknown Question Classification**

A: The question is incorrect because <u>the Wimbledon men's singles trophy does not feature an animal at the top. Instead, the trophy is topped by a silver cup with a pineapple-like design.</u>

Not User-friendly; Fail to Meet User Information Needs

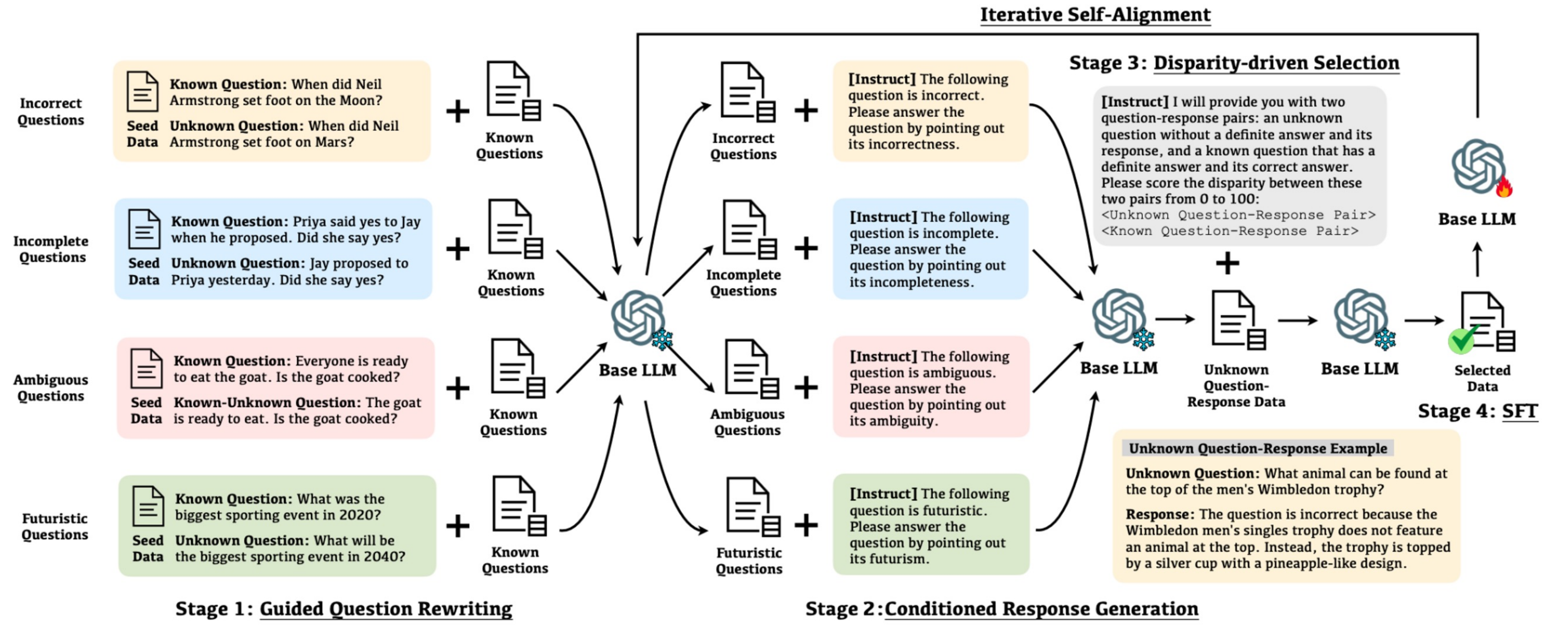
Desired response format:

- ❑ Identify the type of unknown question
- ❑ Provide justifications or explanations

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Workflow of Self-Align

Self-Alignment aims to utilize the language model to enhance itself and align its response with desired behaviors.



Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Initialization

Incorrect Questions

Seed Data
Known Question: When did Neil Armstrong set foot on the Moon?
Unknown Question: When did Neil Armstrong set foot on Mars?

Incomplete Questions

Seed Data
Known Question: Priya said yes to Jay when he proposed. Did she say yes?
Unknown Question: Jay proposed to Priya yesterday. Did she say yes?

Ambiguous Questions

Seed Data
Known Question: Everyone is ready to eat the goat. Is the goat cooked?
Known-Unknown Question: The goat is ready to eat. Is the goat cooked?

Futuristic Questions

Seed Data
Known Question: What was the biggest sporting event in 2020?
Unknown Question: What will be the biggest sporting event in 2040?

Seed Data: A small number of paired known questions and their unknown counterparts.

Base LLM

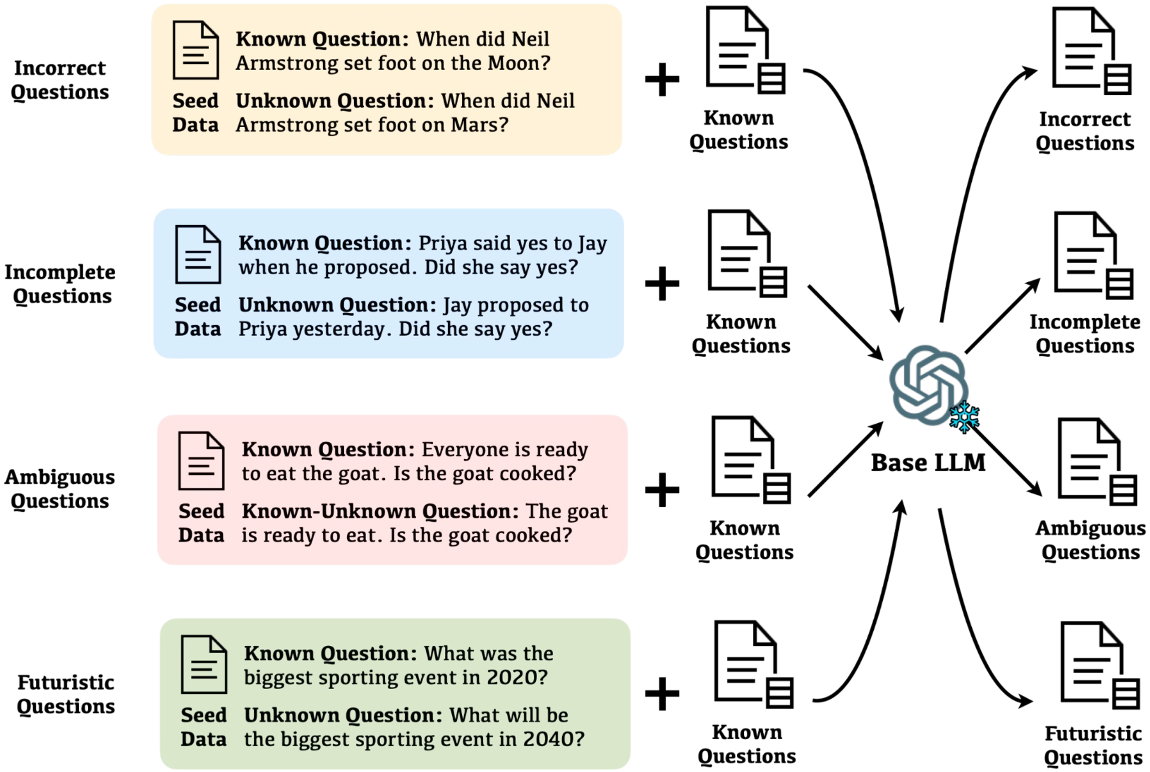
Base LLM: A tunable base LLM to be improved.

Known Questions

Known QA Data: A large number of known question-answer pairs.

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Stage 1: Guided Question Rewriting

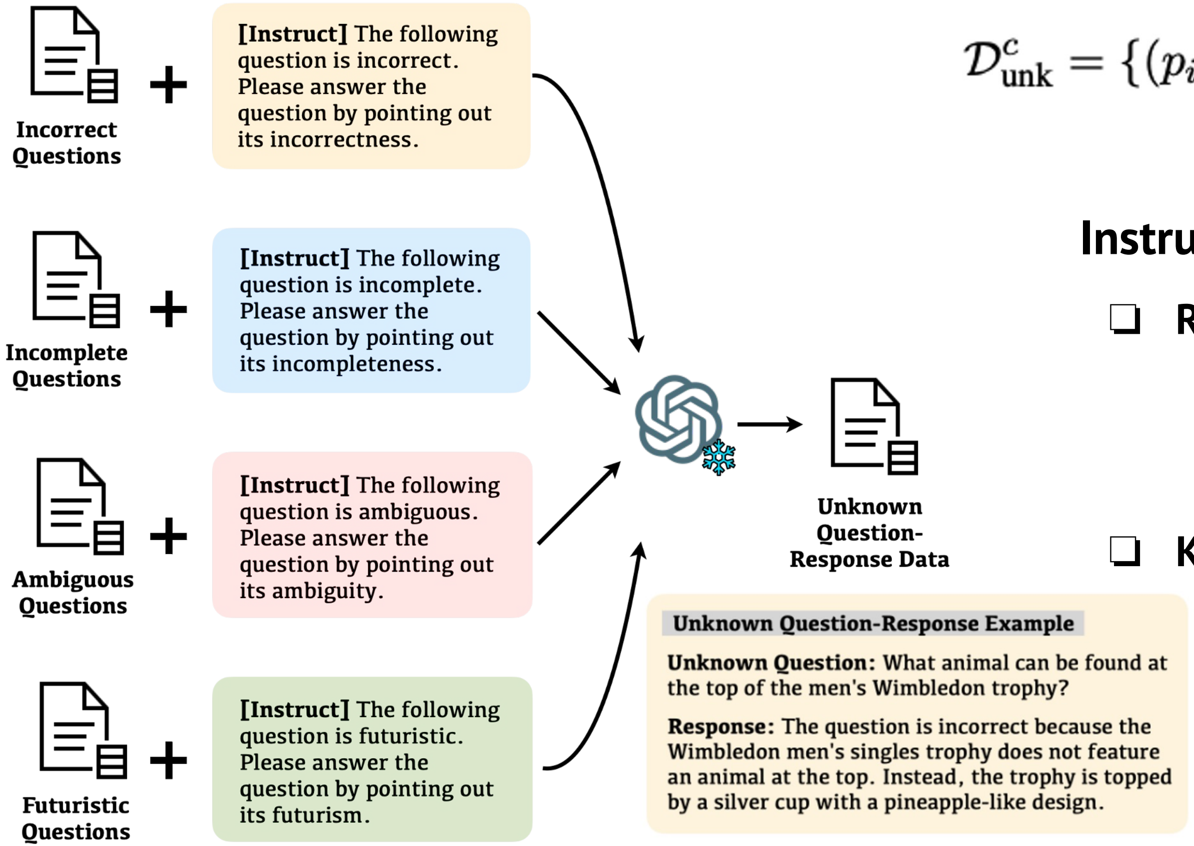


$$\mathcal{D}_{\mathrm{uq}}^{c} = \{\mathcal{M}(z_{qr}^{c}; \mathcal{D}_{\mathrm{seed}}^{c}; q)\}_{q \in \mathcal{D}_{\mathrm{kq}}}$$

- ❑ **Seed Data**
  → demonstrations
- ❑ **Known Questions**
  → source text
- ❑ **Unknown Questions**
  → target text
- ❑ **Base LLM**
  → question rewriter

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Stage 2: Conditioned Response Generation



$$\mathcal{D}^c_{\text{unk}} = \{(p_i, \mathcal{M}(z^c_{rg}; p_i, q_i))\}_{p_i \in \mathcal{D}^c_{\text{uq}}, q_i \in \mathcal{D}^c_{\text{kq}}}$$

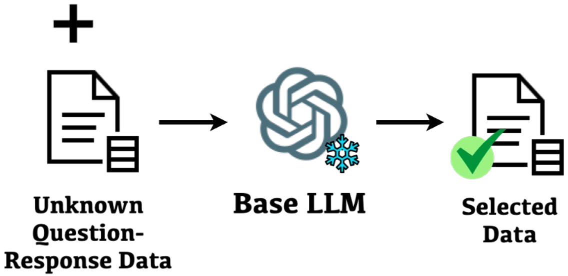
Instructions

- **Response Format**
  - Unknown Question Type
  - Explanation
- **Known Question as Reference**
  - Analyze the unanswerability

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Stage 3: Disparity-driven Self-Curation

[Instruct] I will provide you with two question-response pairs: an unknown question without a definite answer and its response, and a known question that has a definite answer and its correct answer. Please score the disparity between these two pairs from 0 to 100:
`<Unknown Question-Response Pair>`
`<Known Question-Response Pair>`



Unknown Question-Response Example

Unknown Question: What animal can be found at the top of the men's Wimbledon trophy?

Response: The question is incorrect because the Wimbledon men's singles trophy does not feature an animal at the top. Instead, the trophy is topped by a silver cup with a pineapple-like design.

$$s_i = \mathcal{M}(z_{sc}; (q_i, a_i); (p_i, r_i))$$

Why not directly scoring the quality?

➢ The base model itself fails to identify whether the question has a definitive answer.

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Stage 3: Disparity-driven Self-Curation

[Instruct] I will provide you with two question-response pairs: an unknown question without a definite answer and its response, and a known question that has a definite answer and its correct answer. Please score the disparity between these two pairs from 0 to 100:
`<Unknown Question-Response Pair>`
`<Known Question-Response Pair>`

+

Unknown Question-Response Example

Unknown Question: What animal can be found at the top of the men's Wimbledon trophy?

Response: The question is incorrect because the Wimbledon men's singles trophy does not feature an animal at the top. Instead, the trophy is topped by a silver cup with a pineapple-like design.

$$s_i = \mathcal{M}(z_{sc}; (q_i, a_i); (p_i, r_i))$$

Why not directly scoring the quality?

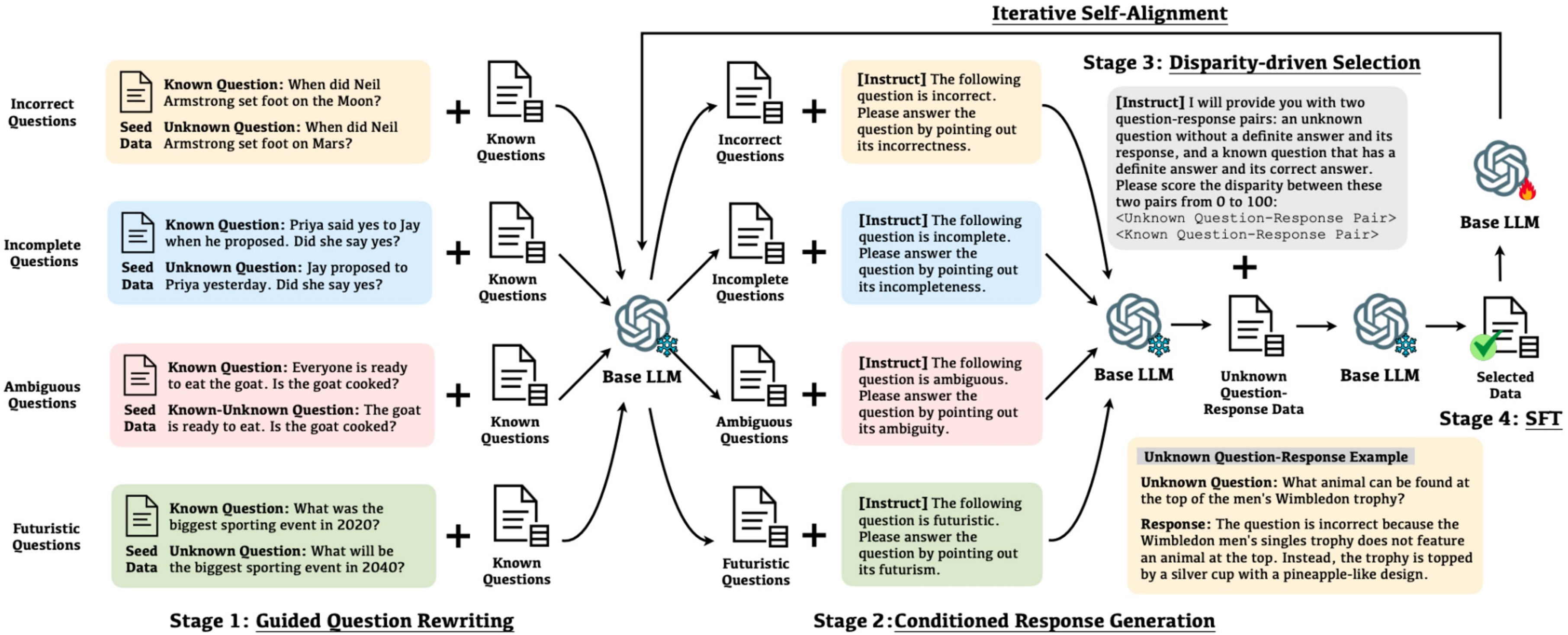
- ➢ The base model itself fails to identify whether the question has a definitive answer.

Why scoring disparity?

- ➢ The conditional generation capability of LLMs ensure the semantic quality of the generated question-response pair.
- ➢ Low disparity score can filter out those low-quality pairs that fail to differentiate from their original known QA counterparts.

Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

Stage 4: Supervised Fine-tuning & Iterative Self-alignment



Deng et al,. "Don't Just Say 'I don't know'! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations" (EMNLP '24)

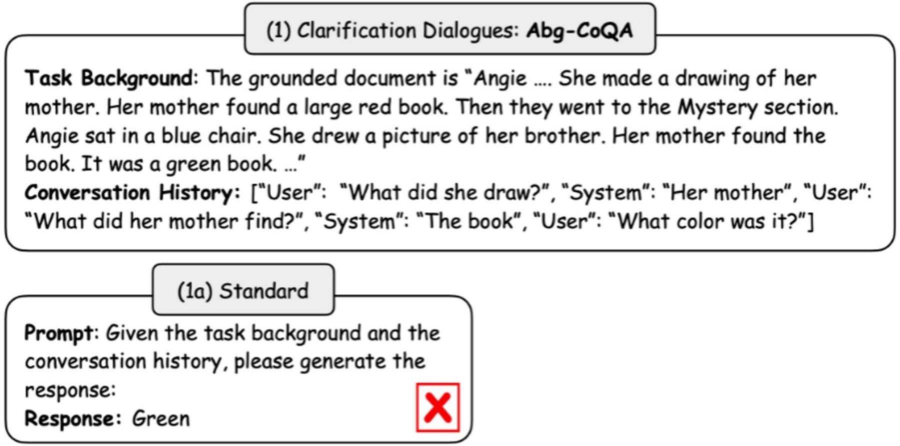
Mitigation of Model-Agnostic Unknown Knowledge

- ❑ **Refusal or Abstention**
    - ❑ Refusal Fine-tuning
    - ❑ Uncertainty-based Reinforcement Learning
    - ❑ Self-alignment
- ❑ **Ask Clarification Questions**
    - ❑ In-Context Learning
    - ❑ Reinforcement Learning
    - ❑ Preference Optimization

Proactive Chain-of-Thought (ProCoT)

- **Standard Prompting**
  - Input: Task Background & Conversation History
  - Output: Response

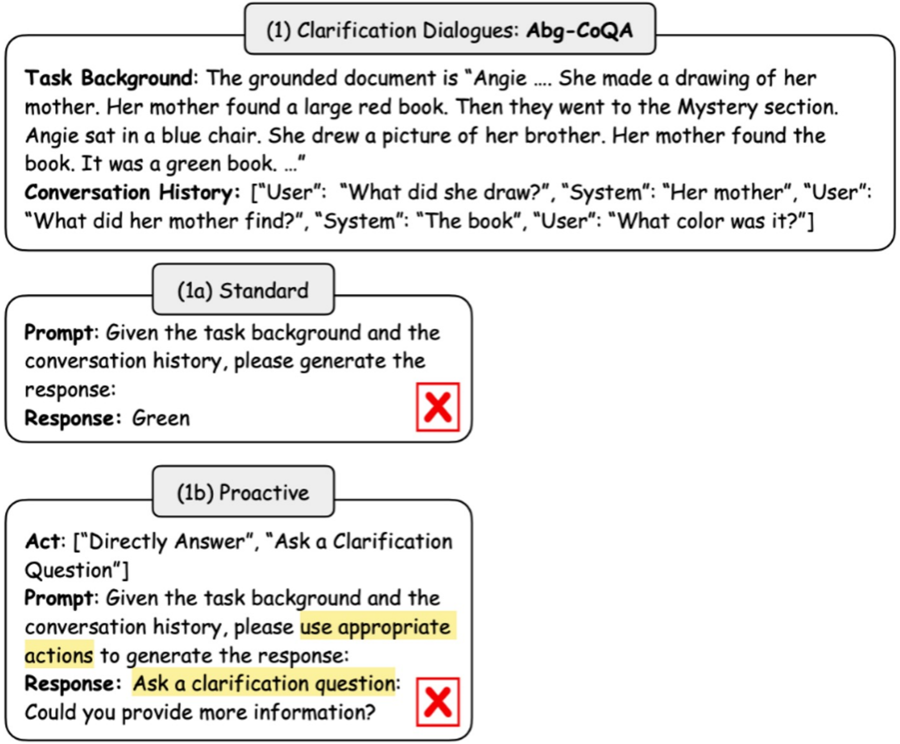
$$p(r|\mathcal{D}, \mathcal{C})$$



Proactive Chain-of-Thought (ProCoT)

- ❑ Standard Prompting
  - ❑ Input: Task Background & Conversation History
  - ❑ Output: Response $p(r|\mathcal{D}, \mathcal{C})$

- ❑ **Proactive Prompting**
  - ❑ Input: + Action Space
  - ❑ Output: + Action $p(a, r|\mathcal{D}, \mathcal{C}, \mathcal{A})$



Deng et al., "Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration" (EMNLP '23 Findings)

Proactive Chain-of-Thought (ProCoT)

- Standard Prompting
  - Input: Task Background & Conversation History
  - Output: Response

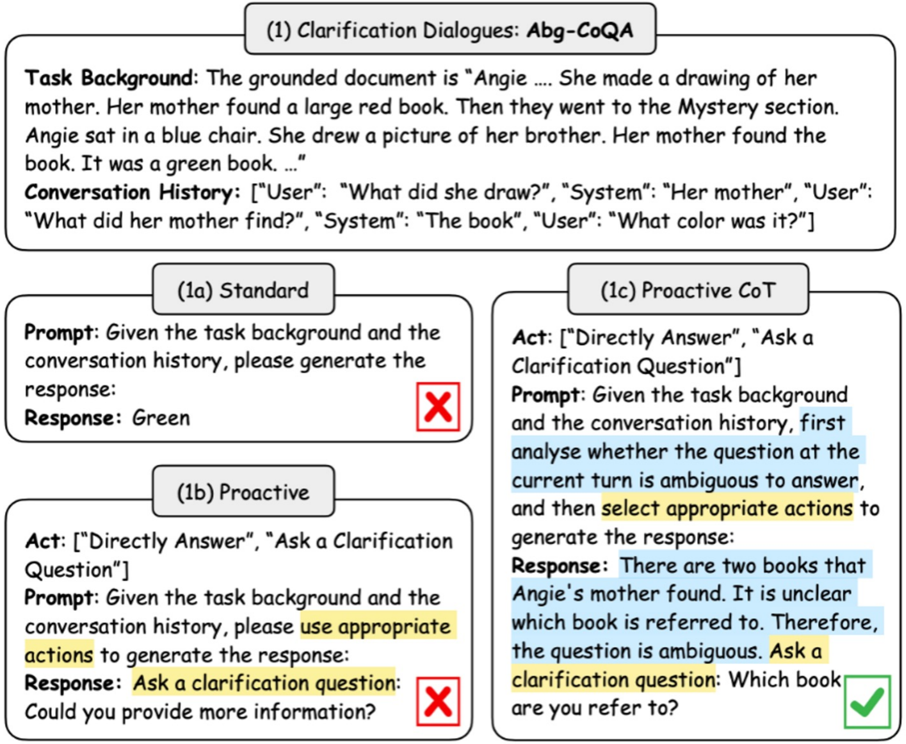
$p(r|\mathcal{D},\mathcal{C})$

- Proactive Prompting
  - Input: + Action Space
  - Output: + Action

$p(a,r|\mathcal{D},\mathcal{C},\mathcal{A})$

- **Proactive Chain-of-Thought Prompting**
  - Output: + Reasoning Chain

$$p(t,a,r|\mathcal{D},\mathcal{C},\mathcal{A})$$



Deng et al., "Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration" (EMNLP '23 Findings)

Proactive Chain-of-Thought (ProCoT)

			Abg-CoQA			PACIFIC		
			CNP	CQG		CNP	CQG	
Method	Shot	Prompt	F1	BLEU-1	Help.	F1	ROUGE-2	Help.
Baseline	-	-	22.1	36.5	30.0	79.0	69.2	38.2
SOTA	-	-	23.6	38.2	56.0	86.9	90.7	80.1
Vicuna-13B	0	Standard	-	11.3	0.0	-	1.2	0.0
	1	Standard	-	11.4	0.0	-	2.5	0.0
	0	Proactive	4.1	13.2	0.0	2.3	2.3	0.0
	1	Proactive	12.1	13.2	4.5	0.0	3.3	0.0
	0	ProCoT	1.4	21.3	9.1	9.7	3.8	10.5
	1	ProCoT	**18.3**	**23.7**	**22.7**	**27.0**	**41.3**	**33.1**
ChatGPT	0	Standard	-	12.1	0.0	-	2.2	0.0
	1	Standard	-	12.3	0.0	-	2.0	0.0
	0	Proactive	22.0	13.7	17.6	19.4	2.9	0.0
	1	Proactive	20.4	**23.4**	23.5	17.7	14.0	12.5
	0	ProCoT	23.8	21.6	32.4	**28.0**	**21.5**	26.7
	1	ProCoT	**27.9**	18.4	**45.9**	27.7	16.2	**35.8**

LLMs barely ask clarification questions, even when the user query is ambiguous.

Deng et al., "Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration" (EMNLP '23 Findings)

Proactive Chain-of-Thought (ProCoT)

			Abg-CoQA (Open-domain)			PACIFIC (Finance)		
			CNP	CQG		CNP	CQG	
Method	Shot	Prompt	F1	BLEU-1	Help.	F1	ROUGE-2	Help.
Baseline	-	-	22.1	36.5	30.0	79.0	69.2	38.2
SOTA	-	-	23.6	38.2	56.0	86.9	90.7	80.1
Vicuna-13B	0	Standard	-	11.3	0.0	-	1.2	0.0
	1	Standard	-	11.4	0.0	-	2.5	0.0
	0	Proactive	4.1	13.2	0.0	2.3	2.3	0.0
	1	Proactive	12.1	13.2	4.5	0.0	3.3	0.0
	0	ProCoT	1.4	21.3	9.1	9.7	3.8	10.5
	1	ProCoT	**18.3**	**23.7**	**22.7**	**27.0**	**41.3**	**33.1**
ChatGPT	0	Standard	-	12.1	0.0	-	2.2	0.0
	1	Standard	-	12.3	0.0	-	2.0	0.0
	0	Proactive	22.0	13.7	17.6	19.4	2.9	0.0
	1	Proactive	20.4	**23.4**	23.5	17.7	14.0	12.5
	0	ProCoT	23.8	21.6	32.4	**28.0**	**21.5**	26.7
	1	ProCoT	**27.9**	18.4	**45.9**	27.7	16.2	**35.8**

LLMs barely ask clarification questions, even when the user query is ambiguous.

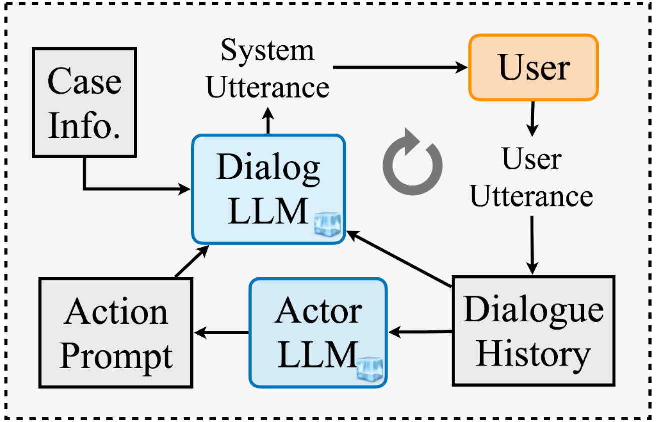
ProCoT largely overcomes this issue in open-domain, but the performance is still unsatisfactory in domain-specific applications.

Deng et al., "Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration" (EMNLP '23 Findings)

Mitigation of Model-Agnostic Unknown Knowledge

- ❑ **Refusal or Abstention**
  - ❑ Refusal Fine-tuning
  - ❑ Uncertainty-based Reinforcement Learning
  - ❑ Self-alignment
- ❑ **Ask Clarification Questions**
  - ❑ In-Context Learning
  - ❑ Reinforcement Learning
  - ❑ Preference Optimization

Limitations of In-context Learning Approaches



- ❑ Fail to optimize the long-term goal of the conversation.
- ❑ Not learnable.
- ❑ Limited by the strategy planning capability of LLMs.

➢ **Reinforcement Learning with Goal-oriented AI Feedback**

Reinforcement Learning

- Formulate the proactive conversation as a **Markov Decision Process (MDP).**
- The objective is to learn a policy π maximizing the expected cumulative rewards over the observed dialogue episodes as:

$$\pi^* = \arg\max_{\pi \in \Pi} \left[\sum_{t=0}^{T} \mathcal{R}(s_t) \right]$$ ***Reward Function***

$$= \arg\max_{\pi \in \Pi} \left[\sum_{t=0}^{T} \mathcal{R}(\mathcal{T}(s_{t-1}, a_t)) \right]$$ ***State Transition***

$$= \arg\max_{\pi \in \Pi} \left[\sum_{t=0}^{T} \mathcal{R}(\mathcal{T}(s_{t-1}, \pi(s_{t-1}))) \right]$$ ***Policy Network***

How to enable the policy learning with LLMs?

Deng et al., "Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents" (ICLR '24)

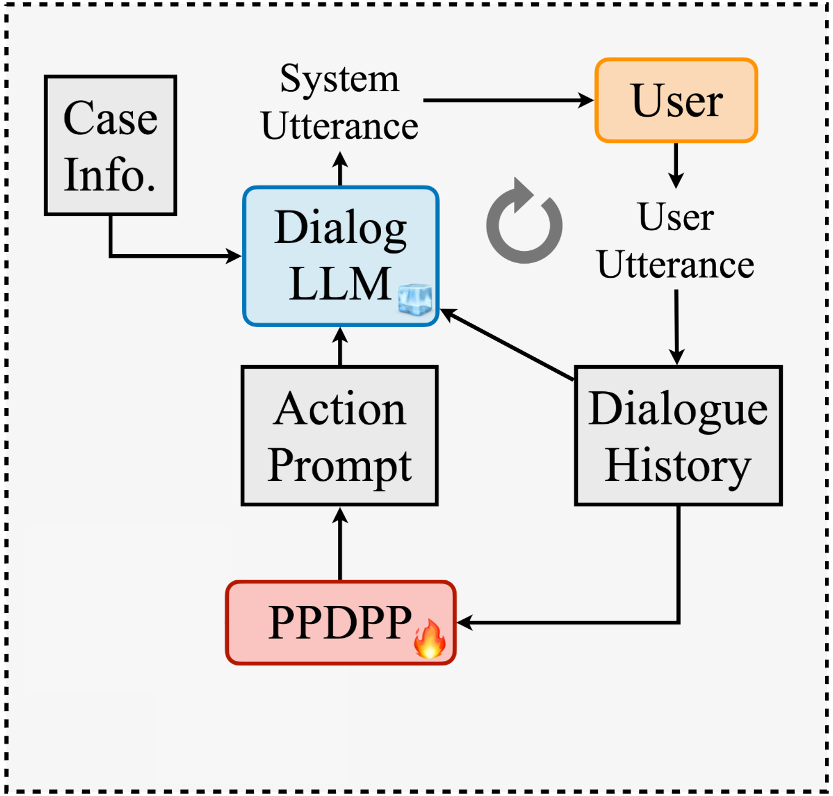
Policy Network – Plug-and-Play Dialogue Policy Planner

- A **tunable language model plug-in** for dialogue strategy learning.

$$a_t = \pi(s_{t-1})$$

- Conduct **Supervised Fine-Tuning** on available human-annotated corpus.

$$\mathcal{L}_c = -\frac{1}{|\mathcal{D}|}\sum\nolimits_{d\in\mathcal{D}}\frac{1}{T_d}\sum\nolimits_{t=1}^{T_d} a_t \log y_t$$



Deng et al., *"Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents" (ICLR '24)*

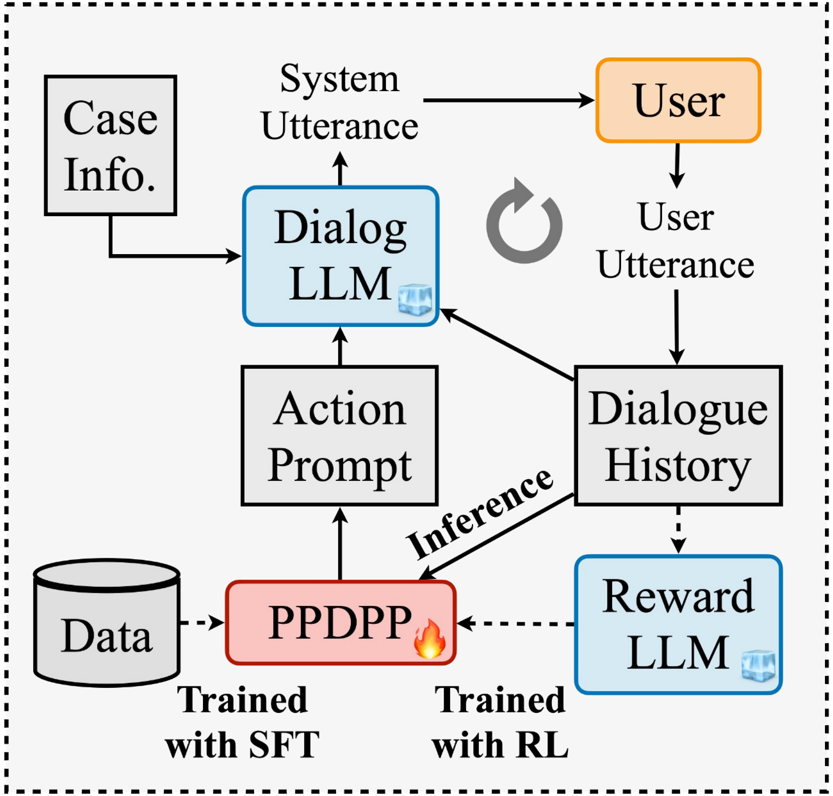
Reward Function – Learning from AI Feedback

- An LLM as the reward model to assess the goal achievement and provide **goal-oriented AI feedback**.

$$\mathcal{R}(s_t) = \frac{1}{l}\sum_{i=1}^{l} \mathcal{M}_r(\mathbf{LLM}_{\mathrm{rwd}}(p_{\mathrm{rwd}}; s_t; \tau))$$

- Employ **Reinforcement Learning** to further tune the policy model.

$$\theta \leftarrow \theta - \alpha \nabla \log \pi_\theta(a_t|s_t) R_t$$



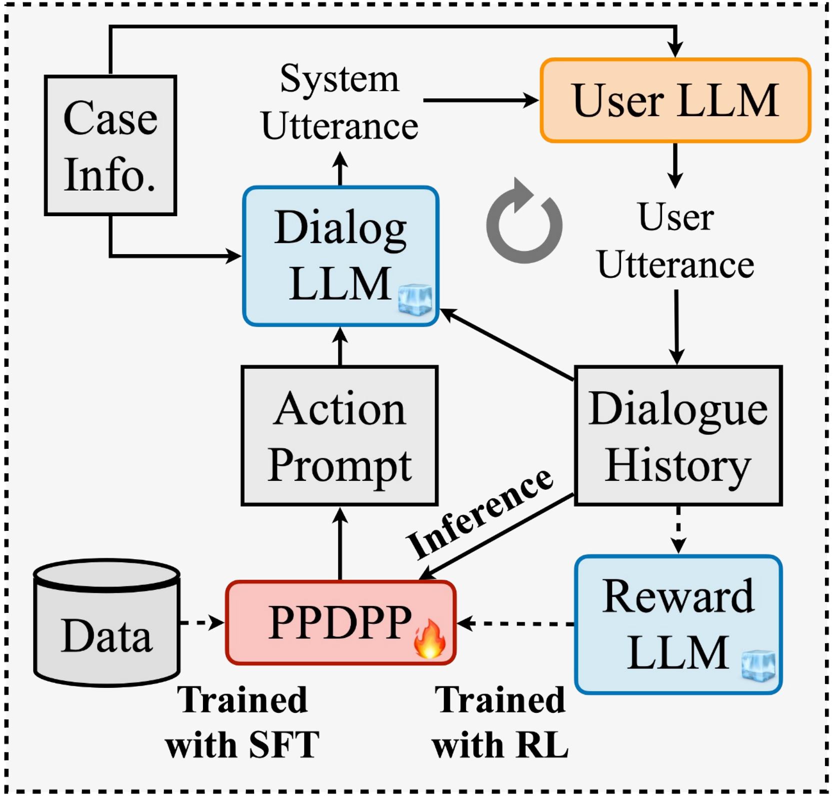
Interacting with real user is costly!

Deng et al., "Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents" (ICLR '24)

State Transition – Multi-agent Simulation

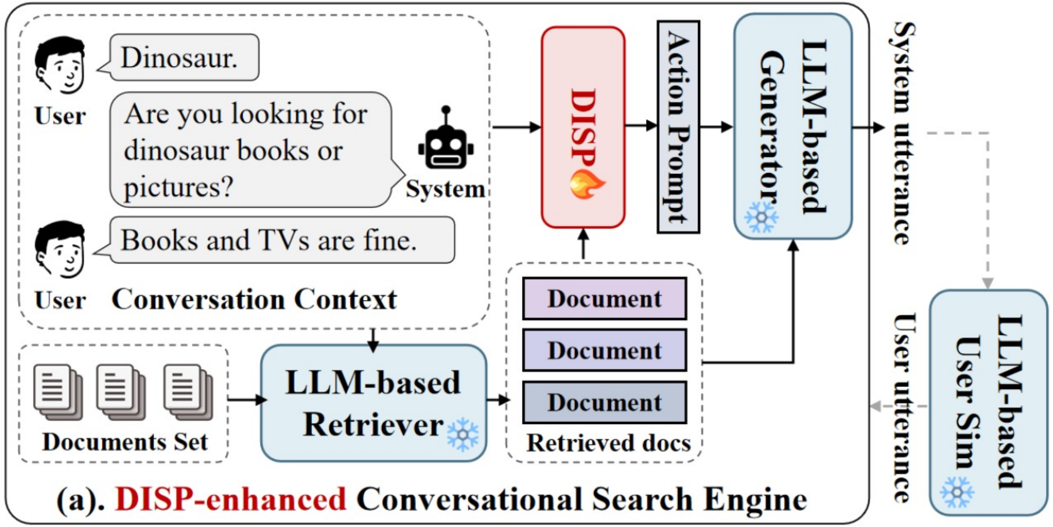
- An LLM to simulate the user with user profiles.
- Employ **Multi-agent Simulation** to collect dynamic interaction data.

$$
\begin{aligned}
u_t^{sys} &= \mathbf{LLM}_{\text{sys}}(p_{\text{sys}}; \mathcal{M}_a(a_t); s_{t-1}) \\
u_t^{usr} &= \mathbf{LLM}_{\text{usr}}(p_{\text{usr}}; s_{t-1}; u_t^{sys}) \\
s_t &= \mathcal{T}(s_{t-1}, a_t) \\
&= \{s_{t-1}; u_t^{sys}, u_t^{usr}\}
\end{aligned}
$$



Deng et al., "Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents" (ICLR '24)

RL for Asking Clarification Questions – STYLE

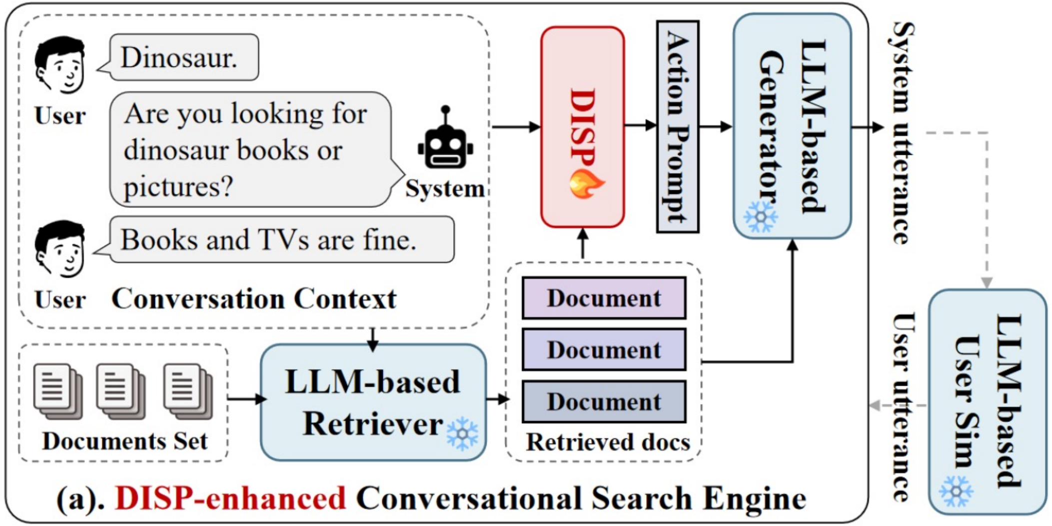


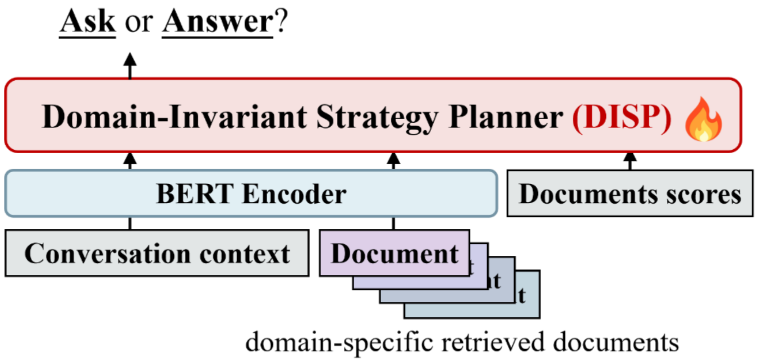
STYLE features rapid transfer to previously unseen domains via tailored strategies.

- ❑ Domain-Invariant Strategy Planner (DISP)
- ❑ Multi-Domain Training (MDT) Paradigm

Chen et al., "STYLE: Improving Domain Transferability of Asking Clarification Questions in Large Language Model Powered Conversational Agents" (ACL '24 Findings)

RL for Asking Clarification Questions – STYLE

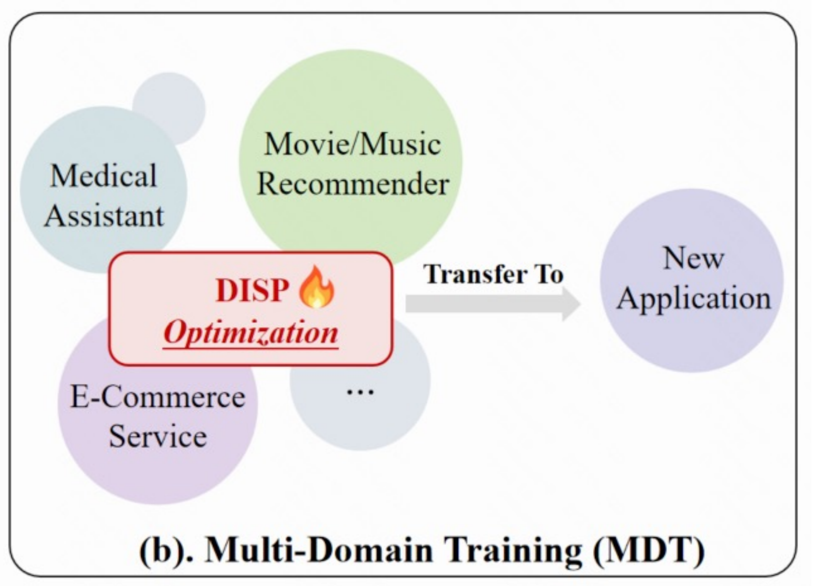




DISP is a policy module that determines when to ask questions. It extract domain-invariant information, mitigating the mismatch in the distribution of domain-specific representations and ensuring robustness across domains.

Chen et al., "STYLE: Improving Domain Transferability of Asking Clarification Questions in Large Language Model Powered Conversational Agents" (ACL '24 Findings)

RL for Asking Clarification Questions – STYLE



$$y_t = \mathbb{E}_{s_{t+1}}\left[r_t + \gamma \max_{a \in \mathcal{A}} Q^*(s_{t+1}, a_{t+1}) \middle| s_t, a_t\right]$$

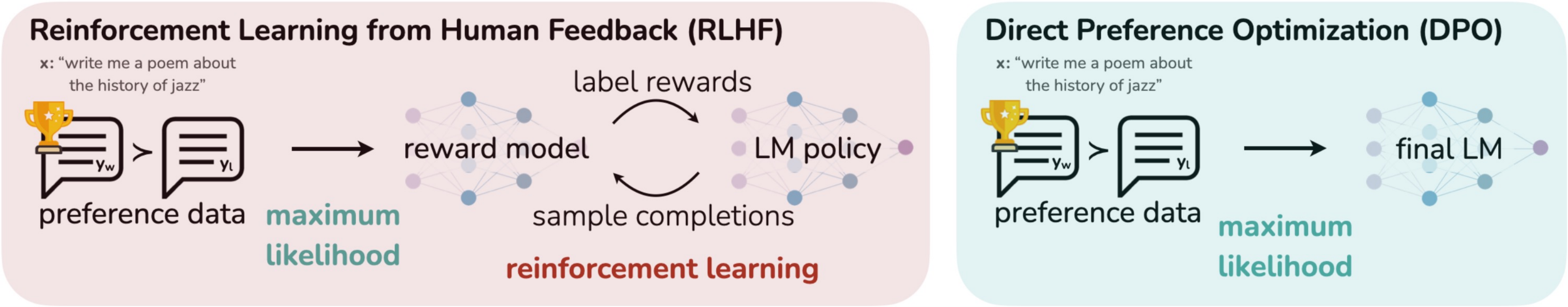
MDT encourages the domain transferability of DISP by training it across multiple diverse domains. This is inspired by the population-based training, which suggests that the generalization of a collaborative agent to held-out populations can be improved by training larger and more diverse populations.

Chen et al., "STYLE: Improving Domain Transferability of Asking Clarification Questions in Large Language Model Powered Conversational Agents" (ACL '24 Findings)

Mitigation of Model-Agnostic Unknown Knowledge

- ❑ **Refusal or Abstention**
  - ❑ Refusal Fine-tuning
  - ❑ Uncertainty-based Reinforcement Learning
  - ❑ Self-alignment
- ❑ **Ask Clarification Questions**
  - ❑ In-Context Learning
  - ❑ Reinforcement Learning
  - ❑ Preference Optimization

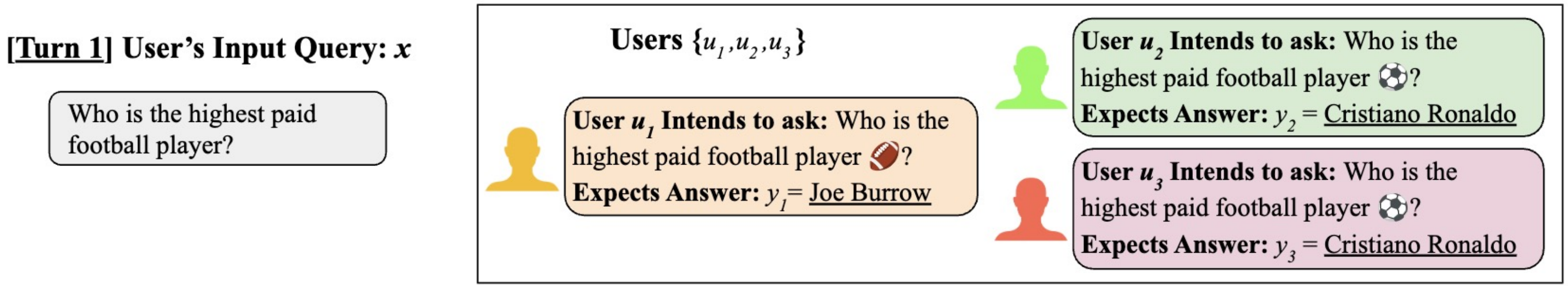
Why Preference Optimization?



- ❑ **No Reward Model Needed**: RLHF/RLAIF requires a separate reward model to be trained on preference data.
- ❑ **No RL Algorithm Needed**: PPO or other RL algorithms could be complex, requiring careful hyperparameter tuning and algorithm designs.
- ❑ **Better Sample Efficiency**: RL requires many environment interactions or sample generations, while DPO operates directly on static preference data.

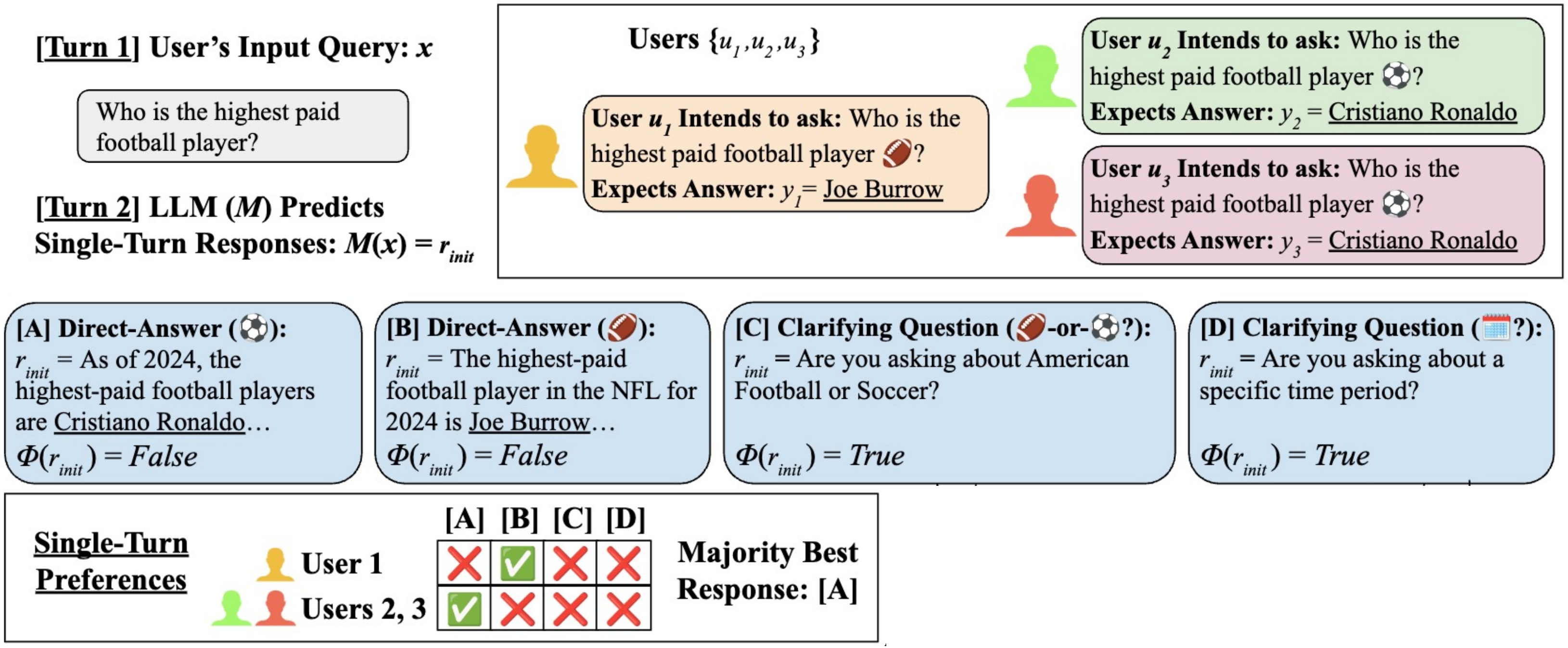
Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model" (NeurIPS '23)

Modeling Future Conversation Turns



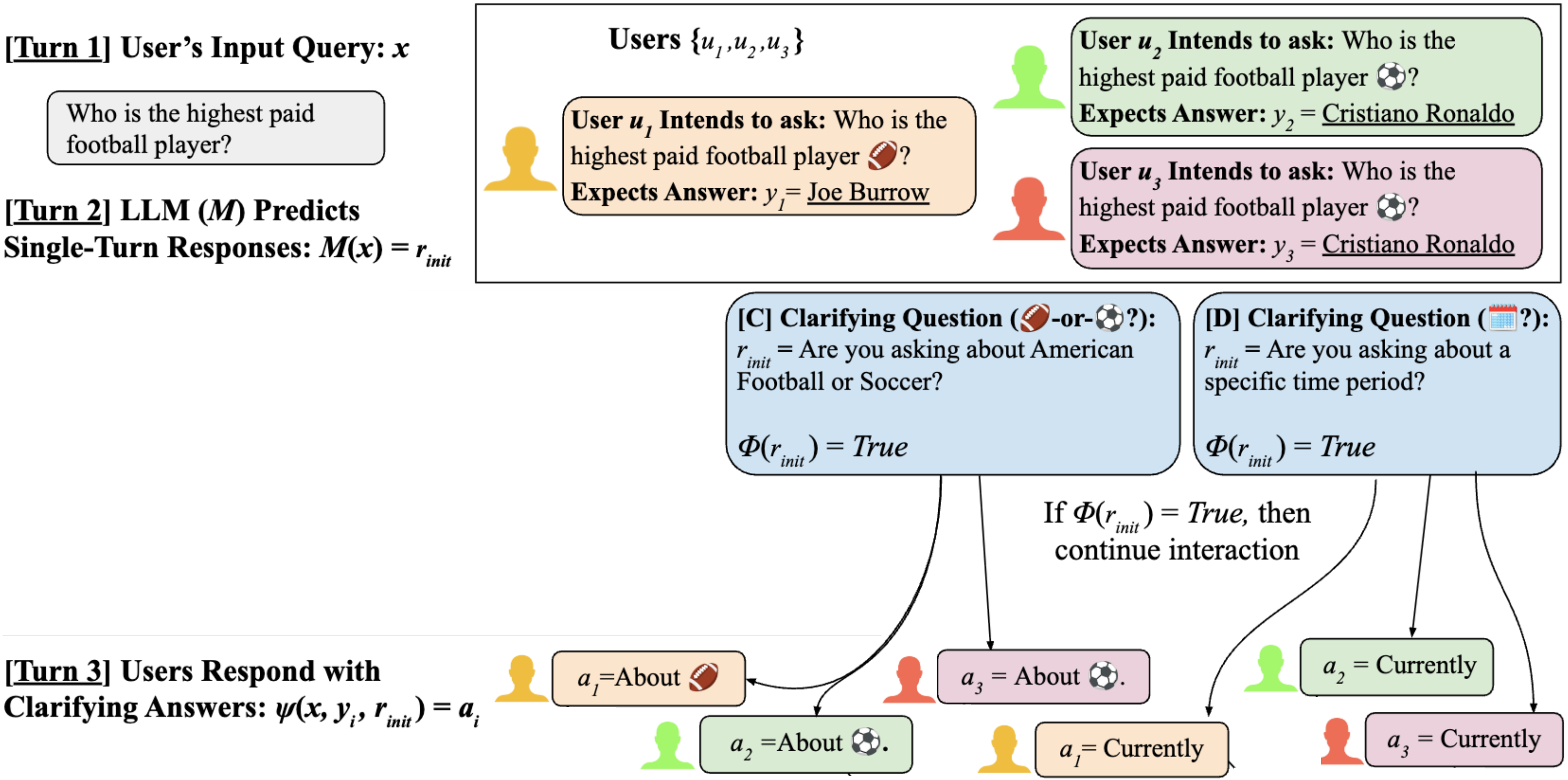
Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

Modeling Future Conversation Turns



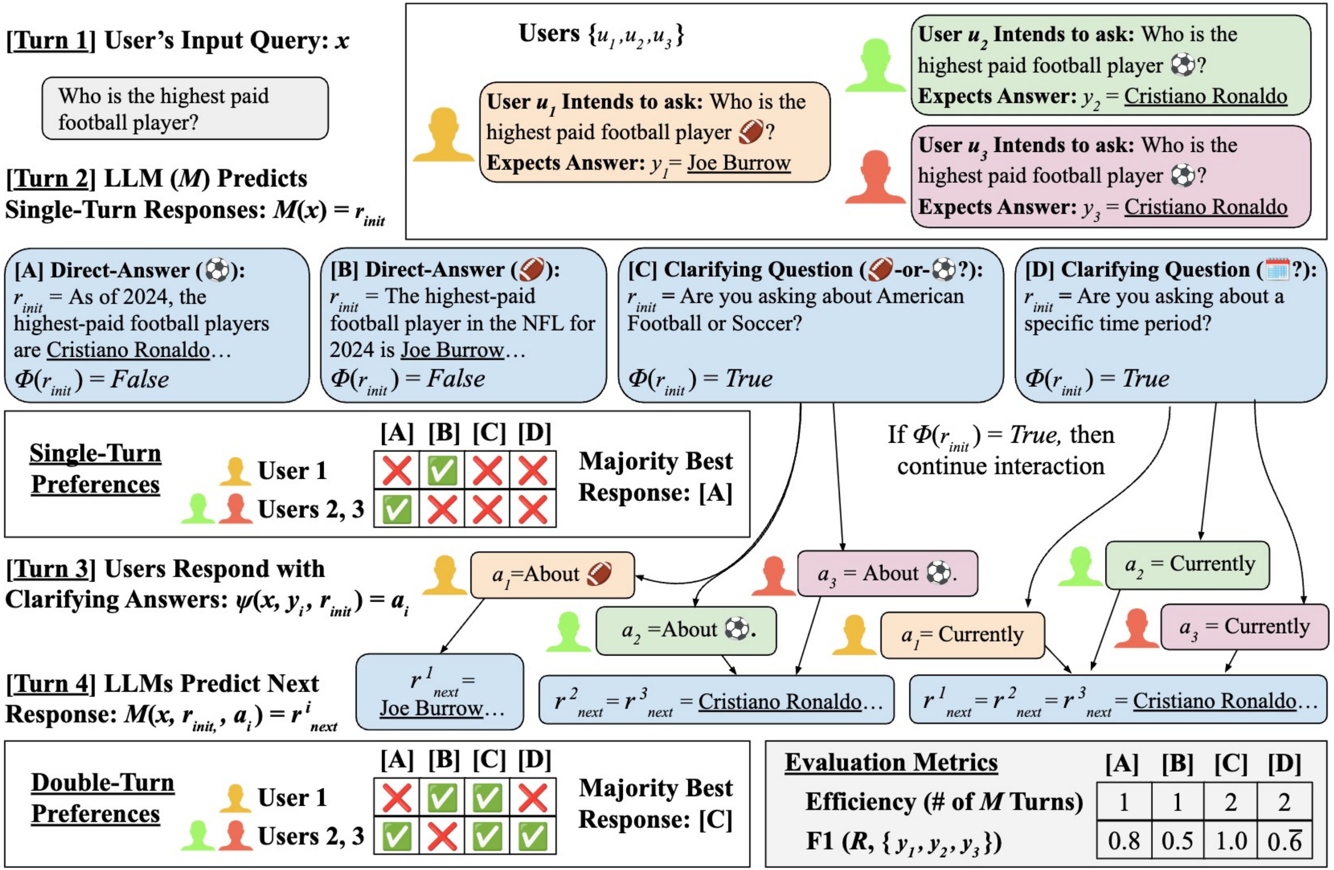
Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

Modeling Future Conversation Turns



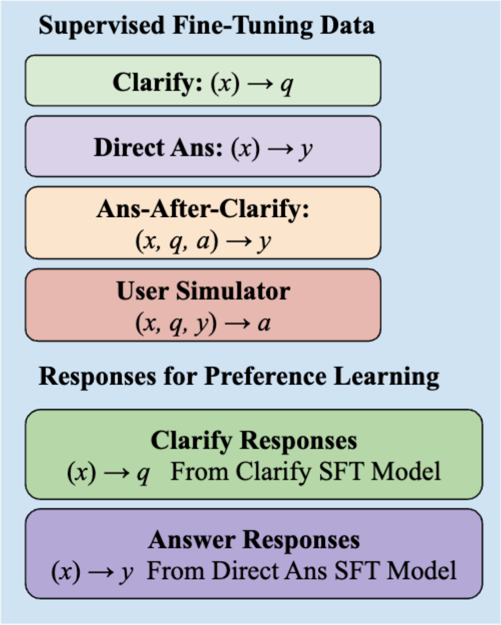
Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

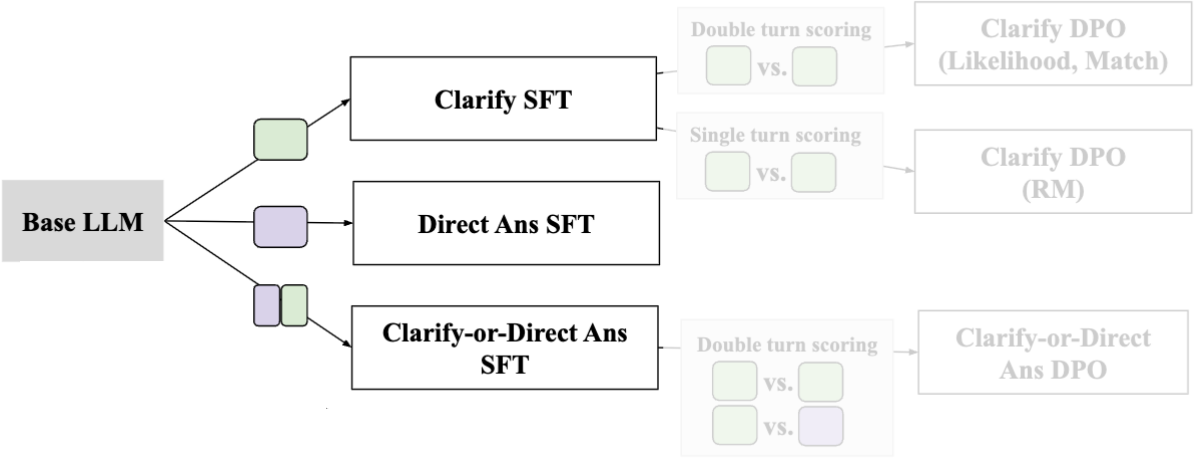
Modeling Future Conversation Turns



Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

Modeling Future Conversation Turns

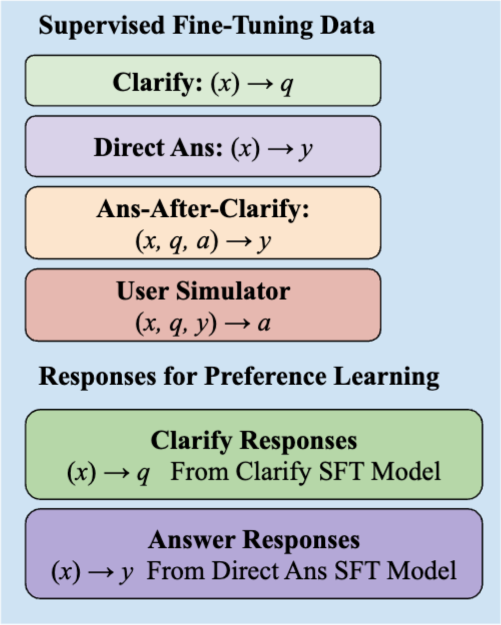


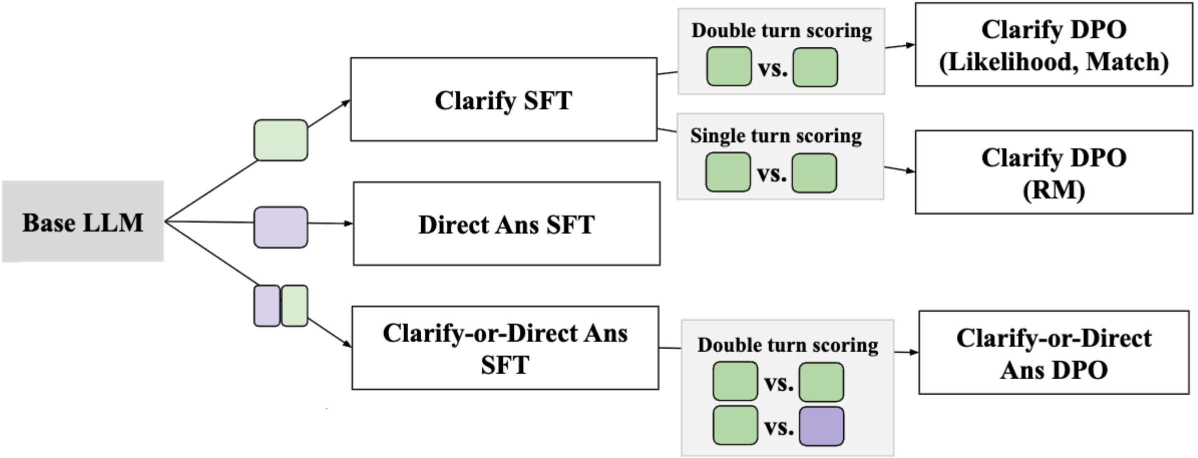


- ❑ **Clarify SFT**: The base LLM is fine-tuned to ask clarifying question to the input query on the SFT data.
- ❑ **Direct-Ans SFT**: The base LLM is fine-tuned on QA data.
- ❑ **Clarify-or-Direct Ans SFT**: The base LLM is fine-tuned on the union of all data used to train Clarify SFT and Direct-Ans SFT models.

Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

Modeling Future Conversation Turns





- ❑ **Clarify DPO**: The Clarify SFT model is further fine-tuned on preference data using DPO.
- ❑ **Clarify-or-Direct Ans DPO**: The Clarify-or-Direct Ans model is further fine-tuned on the ***double-turn preference data*** over clarifying question and direct-answer responses using DPO.

Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

Modeling Future Conversation Turns

	# (↓)	Llama2 Answer F1 (↑) Unamb / Amb / All	# (↓)	Llama3 Answer F1 (↑) Unamb / Amb / All	# (↓)	Gemma Answer F1 (↑) Unamb / Amb / All
Direct-Ans SFT						
w/ Greedy	1	25.4 / 16.8 / 21.1	1	31.2 / 19.2 / 24.8	1	26.1 / 16.8 / 21.1
w/ Sampled	1	25.0 / 17.2 / 21.4	1	28.2 / 20.2 / 24.7	1	23.7 / 17.9 / 21.4
Clarify SFT	2	31.0 / 21.6 / 25.9	2	37.6 / 26.5 / 31.5	2	35.7 / 23.6 / 28.8
Clarify DPO						
w/ RM	2	31.0 / 25.7 / 28.3	2	36.2 / 26.7 / 30.9	2	33.9 / 25.7 / 29.5
w/ Likelihood	2	30.2 / 23.9 / 27.2	2	43.5 / 29.6 / 359	2	37.3 / 26.8 / 31.5
w/ Match	2	**38.3 / 28.2 / 32.8**	2	42.9 / **3.17 / 36.5**	2	**40.7 / 28.6 / 33.9**
Clarify-or-Direct-Ans						
SFT	1.12	25.6 / 18.4 / 21.3	1.40	35.3 / 23.5 / 28.2	1.43	22.3 / 19.0 / 20.3
DPO	1.56	28.9 / 21.1 / 24.3	1.57	35.2 / 25.1 / 29.1	1.61	28.2 / 22.2 / 24.6

Adding a clarifying turn can improve the performance on both ambiguous queries and unambiguous queries.

Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

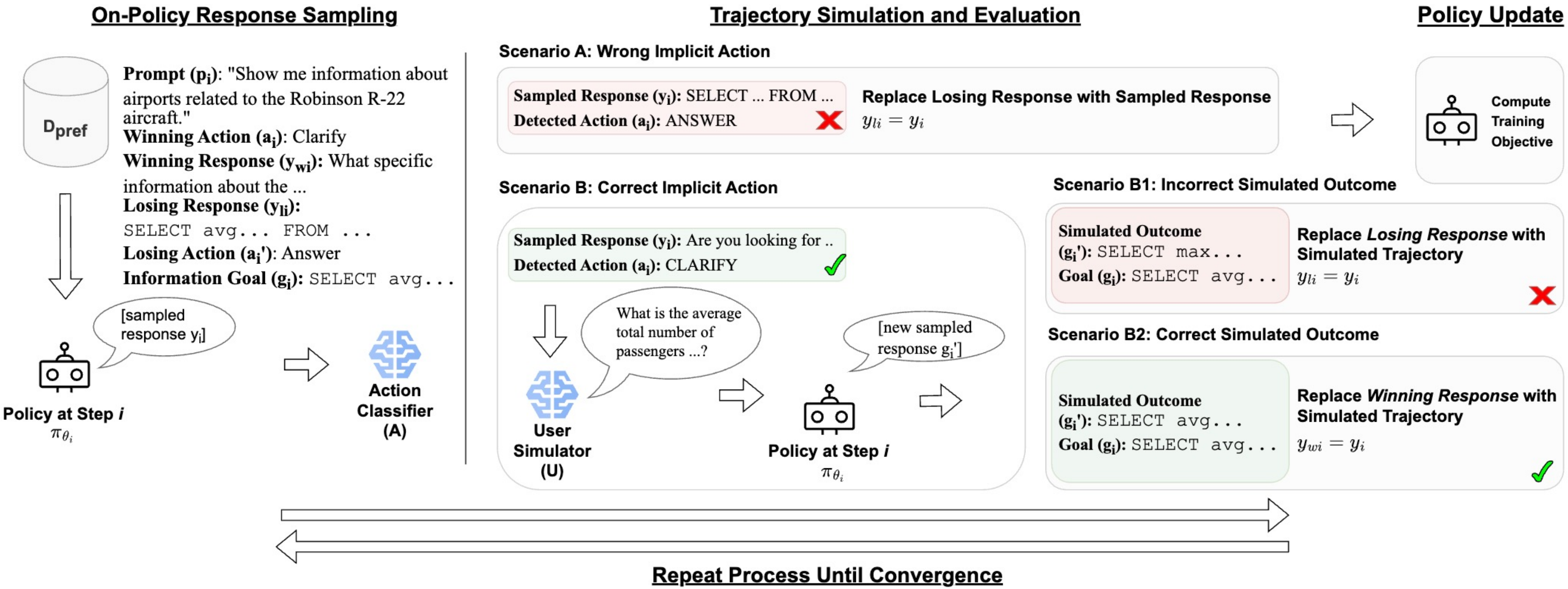
Modeling Future Conversation Turns

	# (↓)	Llama2 Answer F1 (↑) Unamb / Amb / All	# (↓)	Llama3 Answer F1 (↑) Unamb / Amb / All	# (↓)	Gemma Answer F1 (↑) Unamb / Amb / All
Direct-Ans SFT						
w/ Greedy	1	25.4 / 16.8 / 21.1	1	31.2 / 19.2 / 24.8	1	26.1 / 16.8 / 21.1
w/ Sampled	1	25.0 / 17.2 / 21.4	1	28.2 / 20.2 / 24.7	1	23.7 / 17.9 / 21.4
Clarify SFT	2	31.0 / 21.6 / 25.9	2	37.6 / 26.5 / 31.5	2	35.7 / 23.6 / 28.8
Clarify DPO						
w/ RM	2	31.0 / 25.7 / 28.3	2	36.2 / 26.7 / 30.9	2	33.9 / 25.7 / 29.5
w/ Likelihood	2	30.2 / 23.9 / 27.2	2	43.5 / 29.6 / 359	2	37.3 / 26.8 / 31.5
w/ Match	2	**38.3 / 28.2 / 32.8**	2	42.9 / **3.17 / 36.5**	2	**40.7 / 28.6 / 33.9**
Clarify-or-Direct-Ans						
SFT	1.12	25.6 / 18.4 / 21.3	1.40	35.3 / 23.5 / 28.2	1.43	22.3 / 19.0 / 20.3
DPO	1.56	28.9 / 21.1 / 24.3	1.57	35.2 / 25.1 / 29.1	1.61	28.2 / 22.2 / 24.6

- **Clarify-or-Answer methods strike a balance between effectiveness and efficiency.**
- **DPO with double-turn preference data consistently outperforms SFT.**

Zhang et al., "Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions" (ICLR '25)

Action-Based Contrastive Self-Training (ACT)



- ❑ **ACT** focuses on the clarification preference optimization in multi-turn conversations
- ❑ Construct conversation data with contrastive action pairs (*clarify* or *answer*) as the preference data

Chen et al., "Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training" (ICLR '25)