Learning to Clarify: Multi-turn Conversations Through Action-Based Contrastive Self-Training

Maximillian Chen, Ruoxi Sun, Tomas Pfister, Sercan O. Arik
Google & Columbia University
ICLR 2025 Paper

Google Cloud

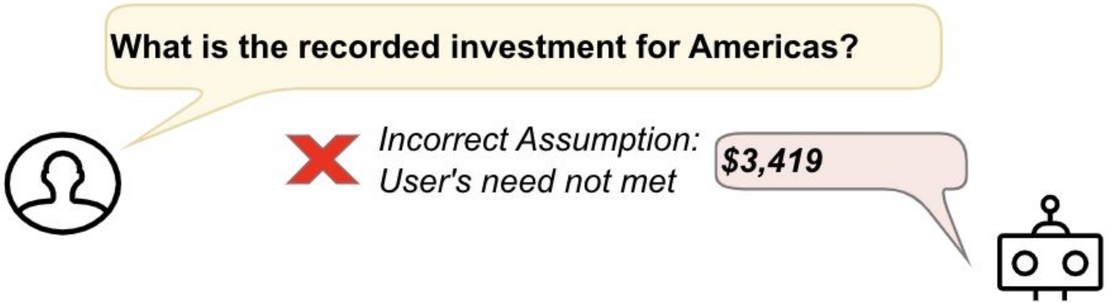
- Conversations are filled with ambiguity – especially in online interaction

Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395

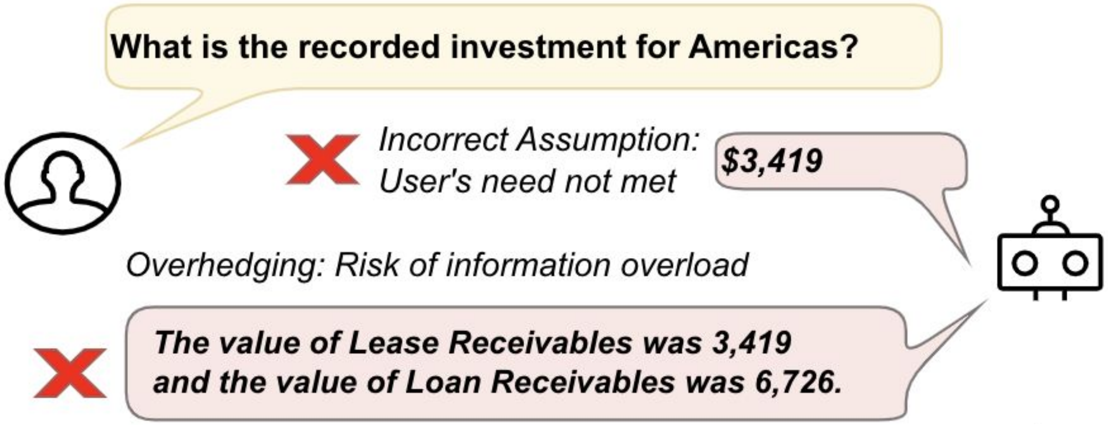
What is the recorded investment for Americas?

- Conversations are filled with ambiguity – especially in online interaction

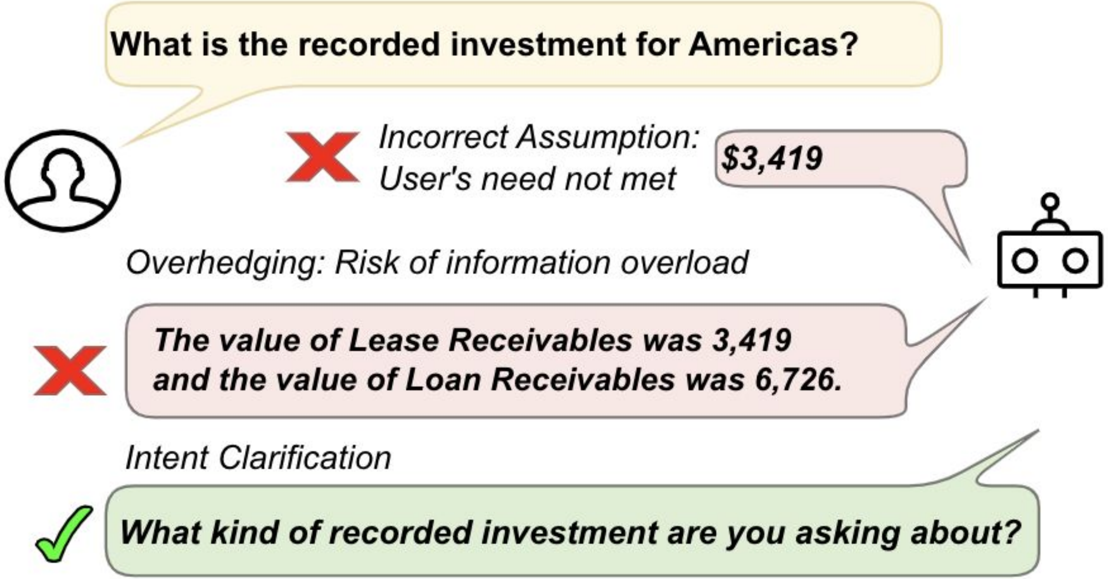
Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



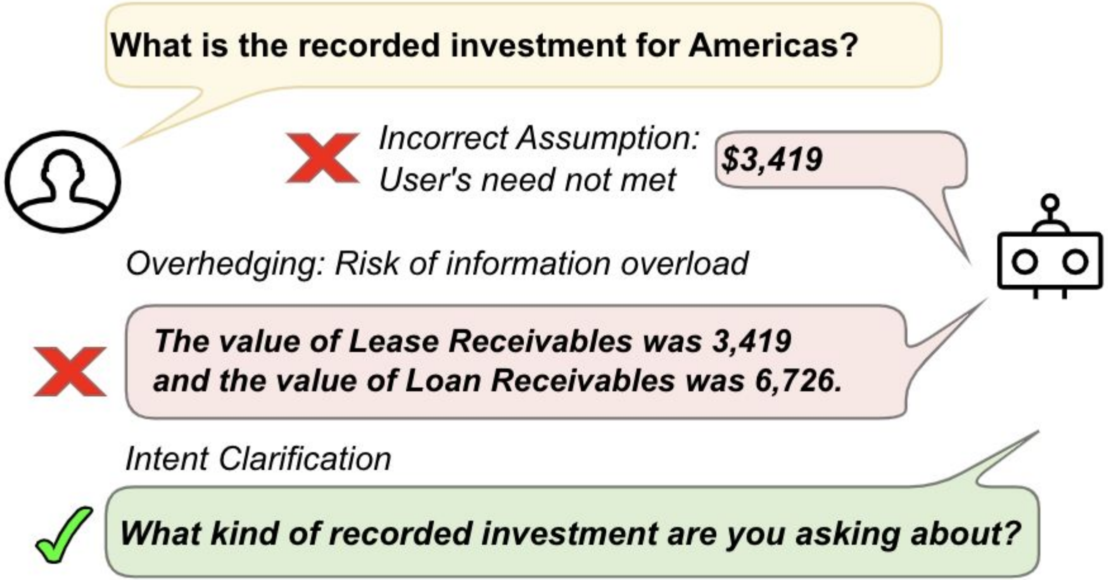
Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



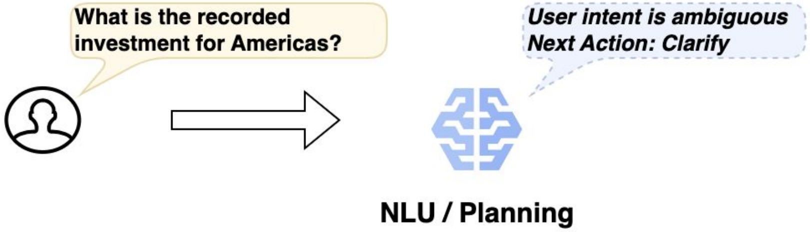
- Conversations are “mixed-initiative”
- Assistants should know how to collaborate with users to achieve goal outcomes in various complex settings, rather than one-way QA

What can we learn from traditional dialogue system approaches?

Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395

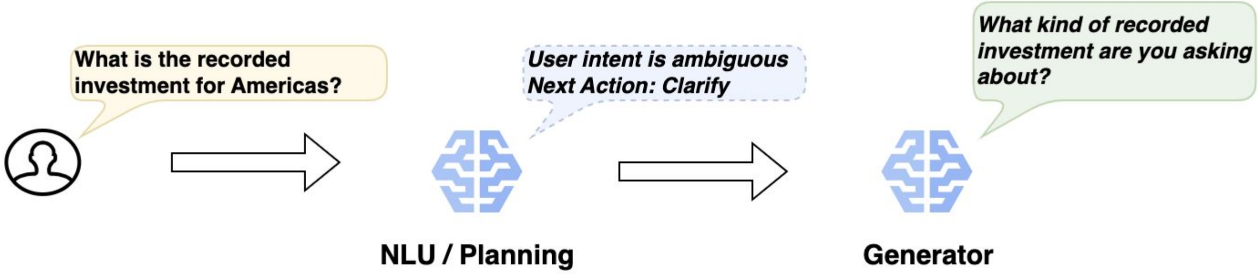
What can we learn from traditional dialogue system approaches?

Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



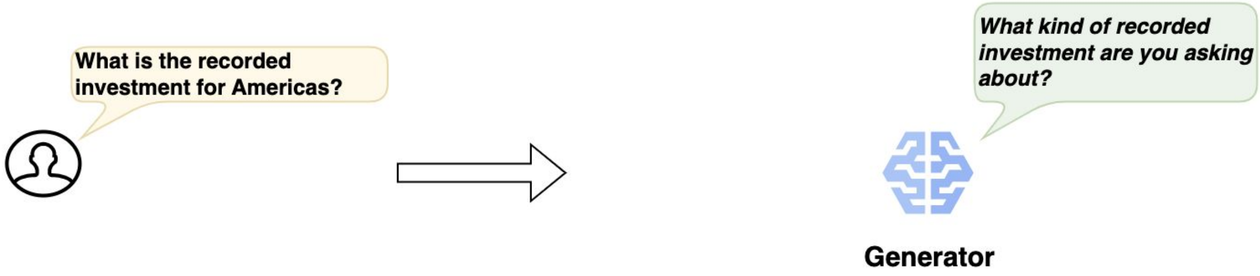
What can we learn from traditional dialogue system approaches?

Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



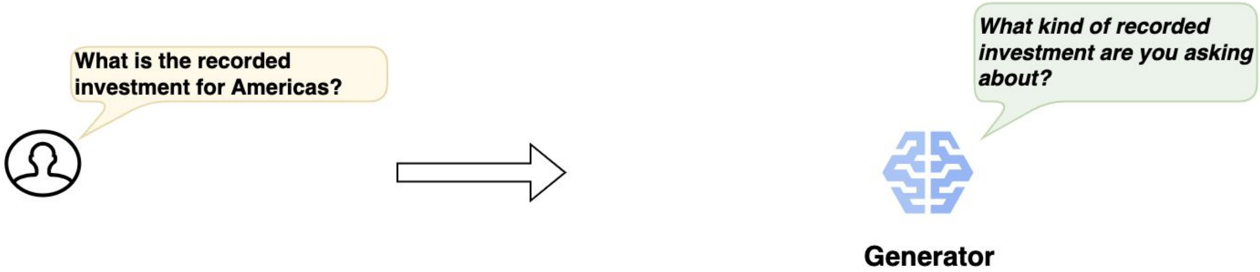
LLM-based assistants skip planning, but take no measures to optimize generation with planning in mind.

	Recorded Investments		
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



LLM-based assistants skip planning, but take no measures to optimize generation with planning in mind.

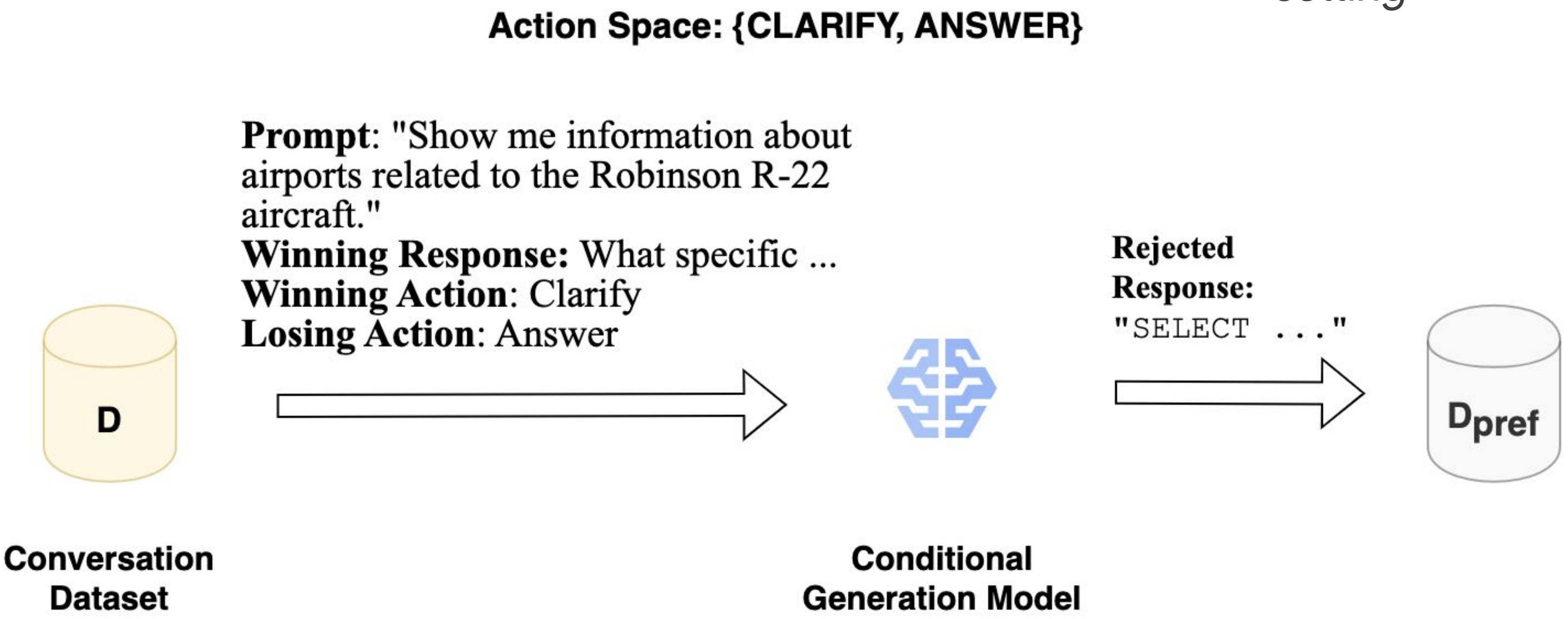
Recorded Investments			
	Americas	EMEA	Asia Pacific
Lease Receivables	3,419	1,186	963
Loan Receivables	6,726	3,901	2,395



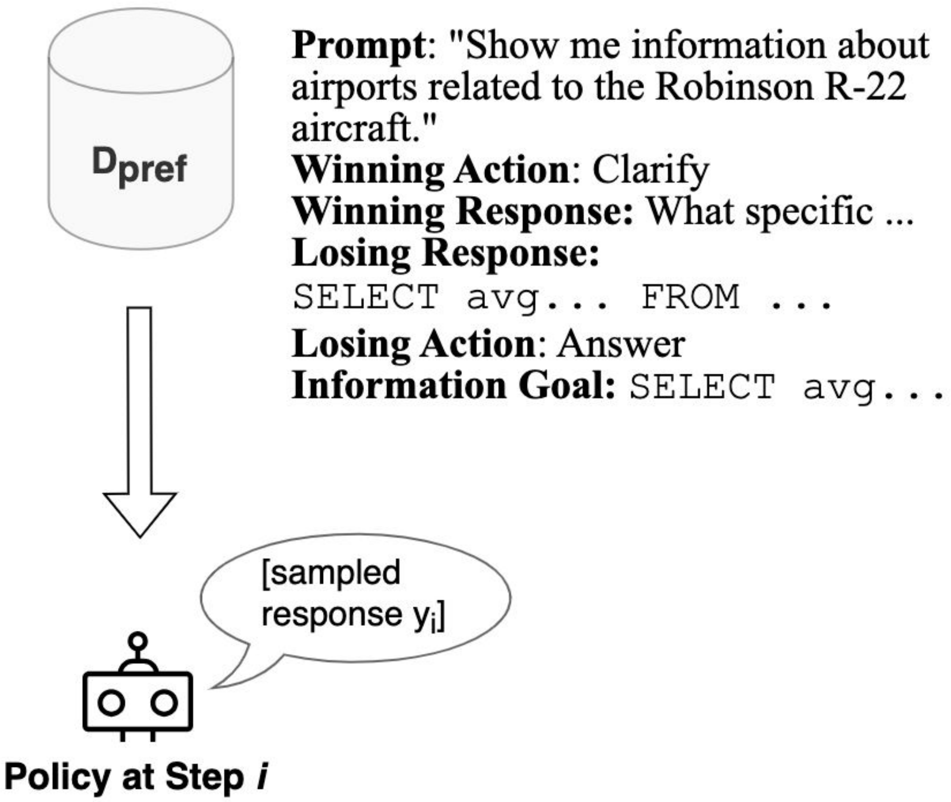
…if we keep this interaction paradigm in mind, can we then <u>learn planning implicitly?</u>

We propose ACT: Action-Based Contrastive Self-Training

Contrastive pairings make sense for our setting

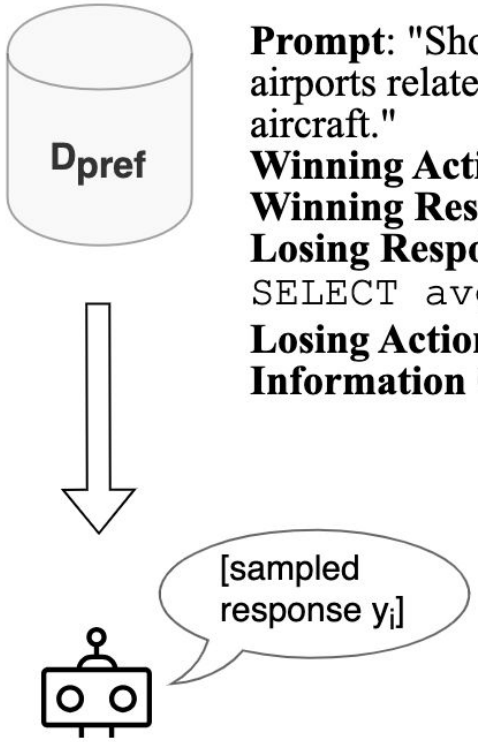


On-Policy Response Sampling



A sampled response will likely have a higher log probability than any response from a static dataset or an outside LLM

On-Policy Response Sampling



Prompt: "Show me information about airports related to the Robinson R-22 aircraft."
Winning Action: Clarify
Winning Response: What specific ...
Losing Response:
`SELECT avg... FROM ...`
Losing Action: Answer
Information Goal: `SELECT avg...`

Policy at Step *i*

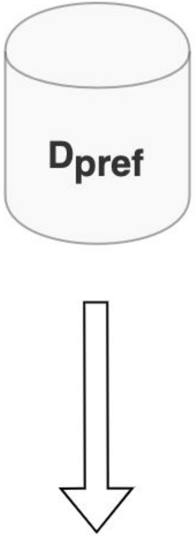
Action Classifier

Scenario A: Wrong Implicit Action

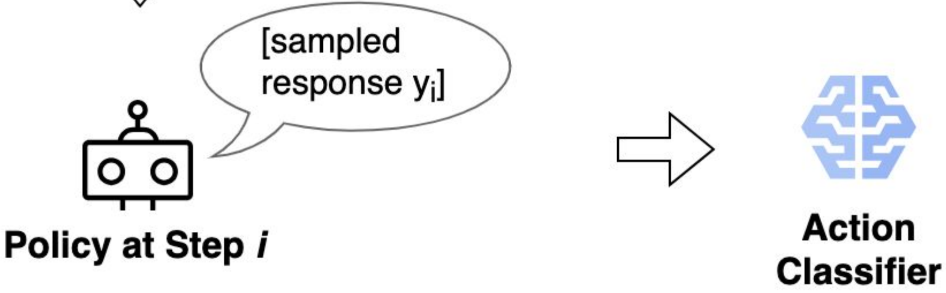
Sampled Response: SELECT ... FROM ...
Detected Action: ANSWER ❌

Replace Losing Response with Sampled Response

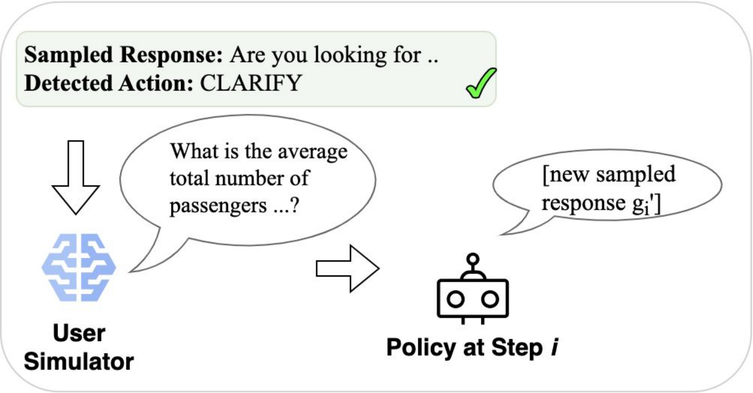
On-Policy Response Sampling

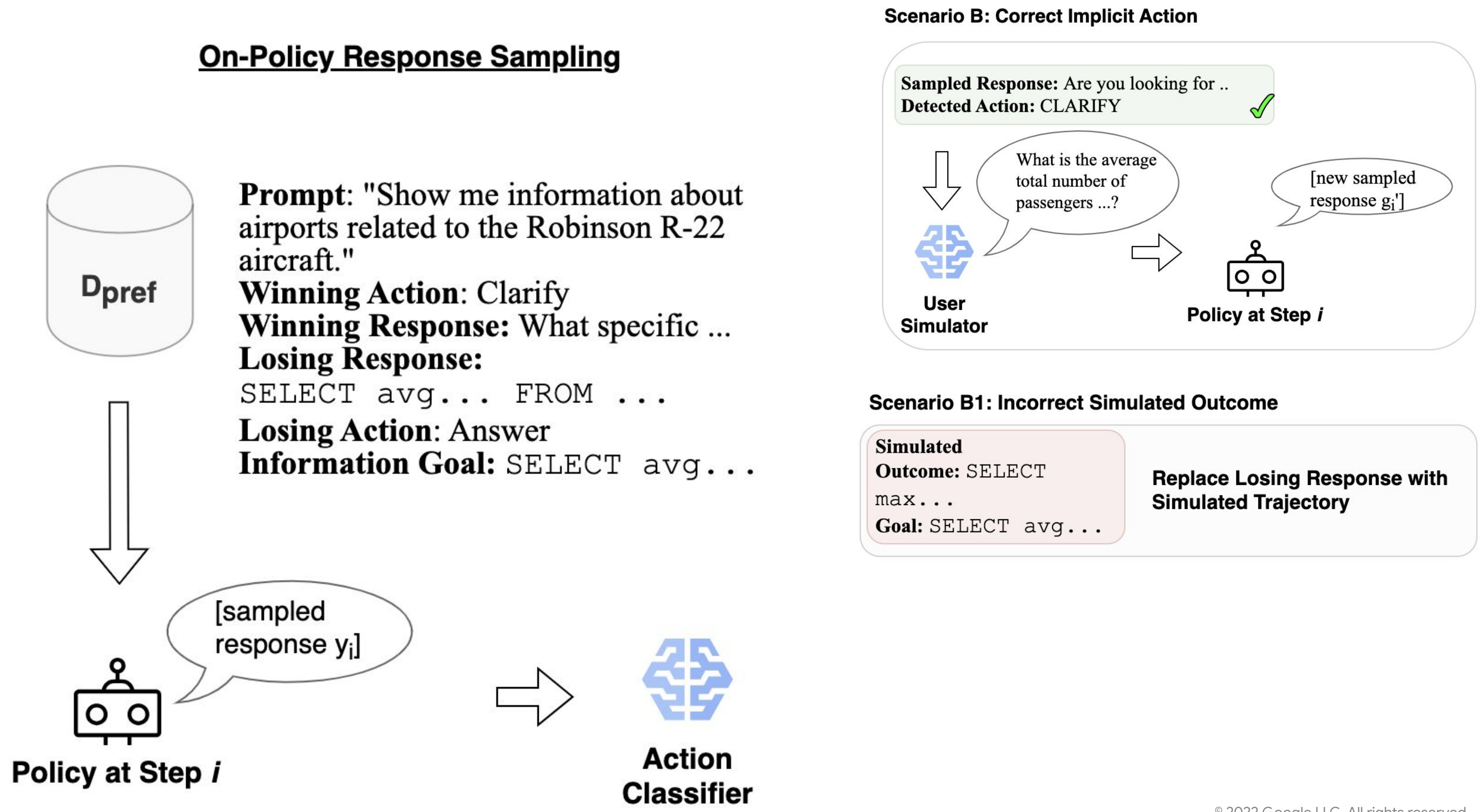


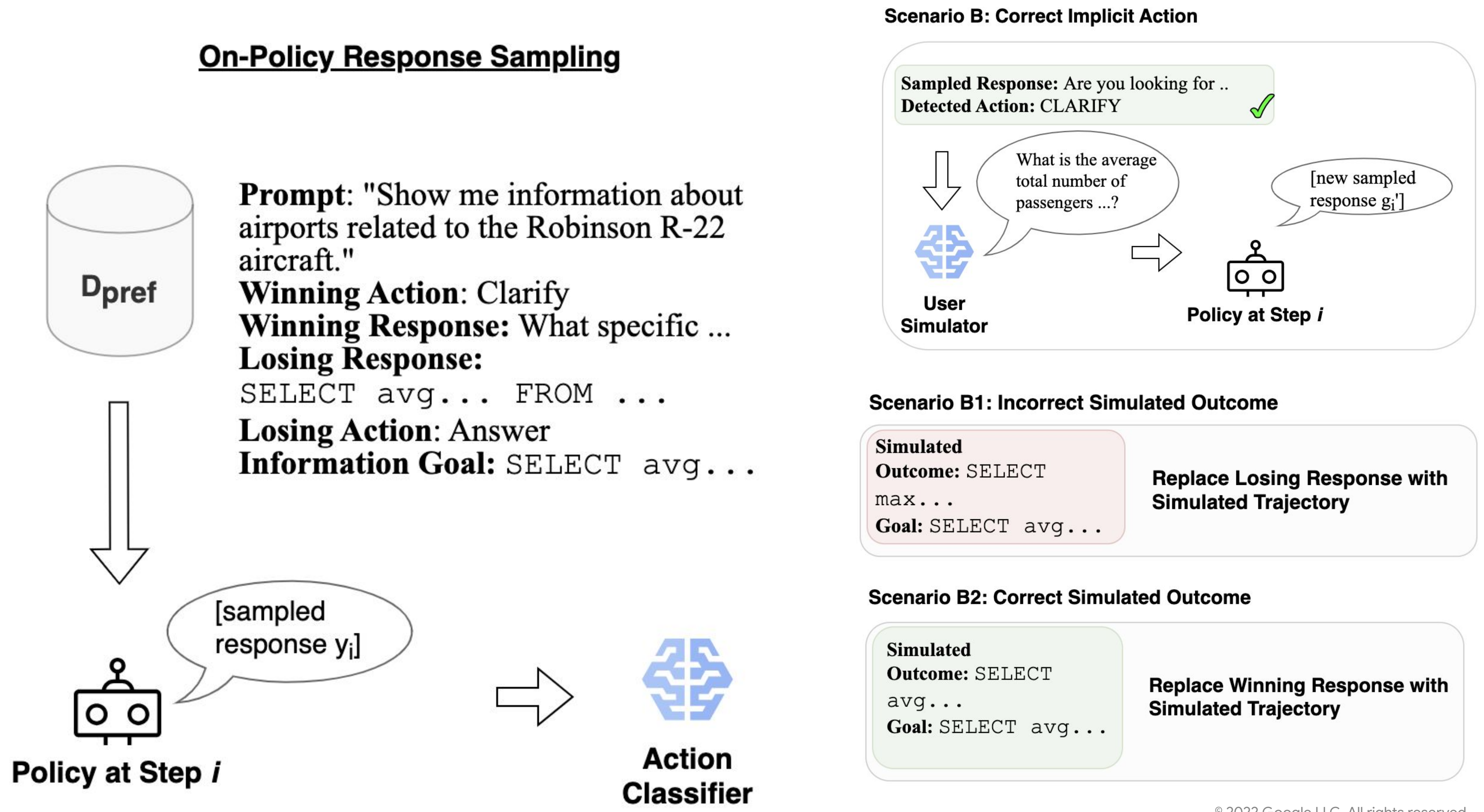
Prompt: "Show me information about airports related to the Robinson R-22 aircraft."
Winning Action: Clarify
Winning Response: What specific ...
Losing Response:
`SELECT avg... FROM ...`
Losing Action: Answer
Information Goal: `SELECT avg...`



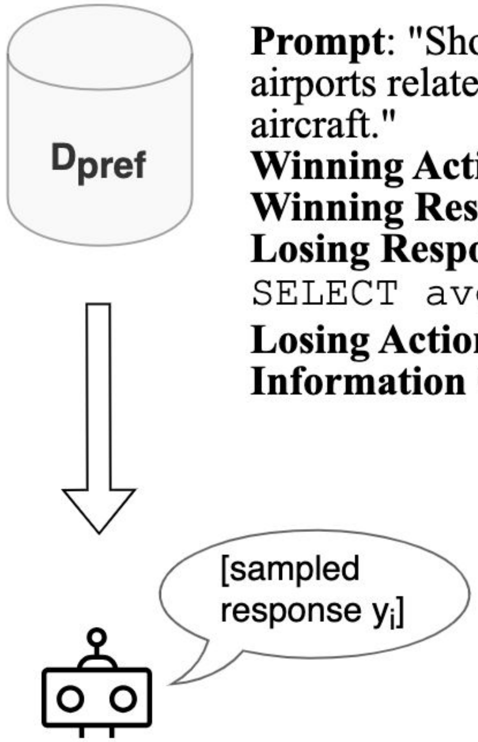
Scenario B: Correct Implicit Action




On-Policy Response Sampling
Dpref
Prompt: "Show me information about airports related to the Robinson R-22 aircraft."
Winning Action: Clarify
Winning Response: What specific ...
Losing Response:
SELECT avg... FROM ...
Losing Action: Answer
Information Goal: SELECT avg...
[sampled response yi]
Policy at Step i
Action Classifier
Scenario B: Correct Implicit Action
Sampled Response: Are you looking for ..
Detected Action: CLARIFY
What is the average total number of passengers ...?
[new sampled response gi']
User Simulator
Policy at Step i
Scenario B1: Incorrect Simulated Outcome
Simulated Outcome: SELECT max...
Goal: SELECT avg...
Replace Losing Response with Simulated Trajectory


On-Policy Response Sampling
Dpref
Prompt: "Show me information about airports related to the Robinson R-22 aircraft."
Winning Action: Clarify
Winning Response: What specific ...
Losing Response:
SELECT avg... FROM ...
Losing Action: Answer
Information Goal: SELECT avg...
[sampled response yi]
Policy at Step i
Action Classifier
Scenario B: Correct Implicit Action
Sampled Response: Are you looking for ..
Detected Action: CLARIFY
What is the average total number of passengers ...?
[new sampled response gi']
User Simulator
Policy at Step i
Scenario B1: Incorrect Simulated Outcome
Simulated Outcome: SELECT max...
Goal: SELECT avg...
Replace Losing Response with Simulated Trajectory
Scenario B2: Correct Simulated Outcome
Simulated Outcome: SELECT avg...
Goal: SELECT avg...
Replace Winning Response with Simulated Trajectory

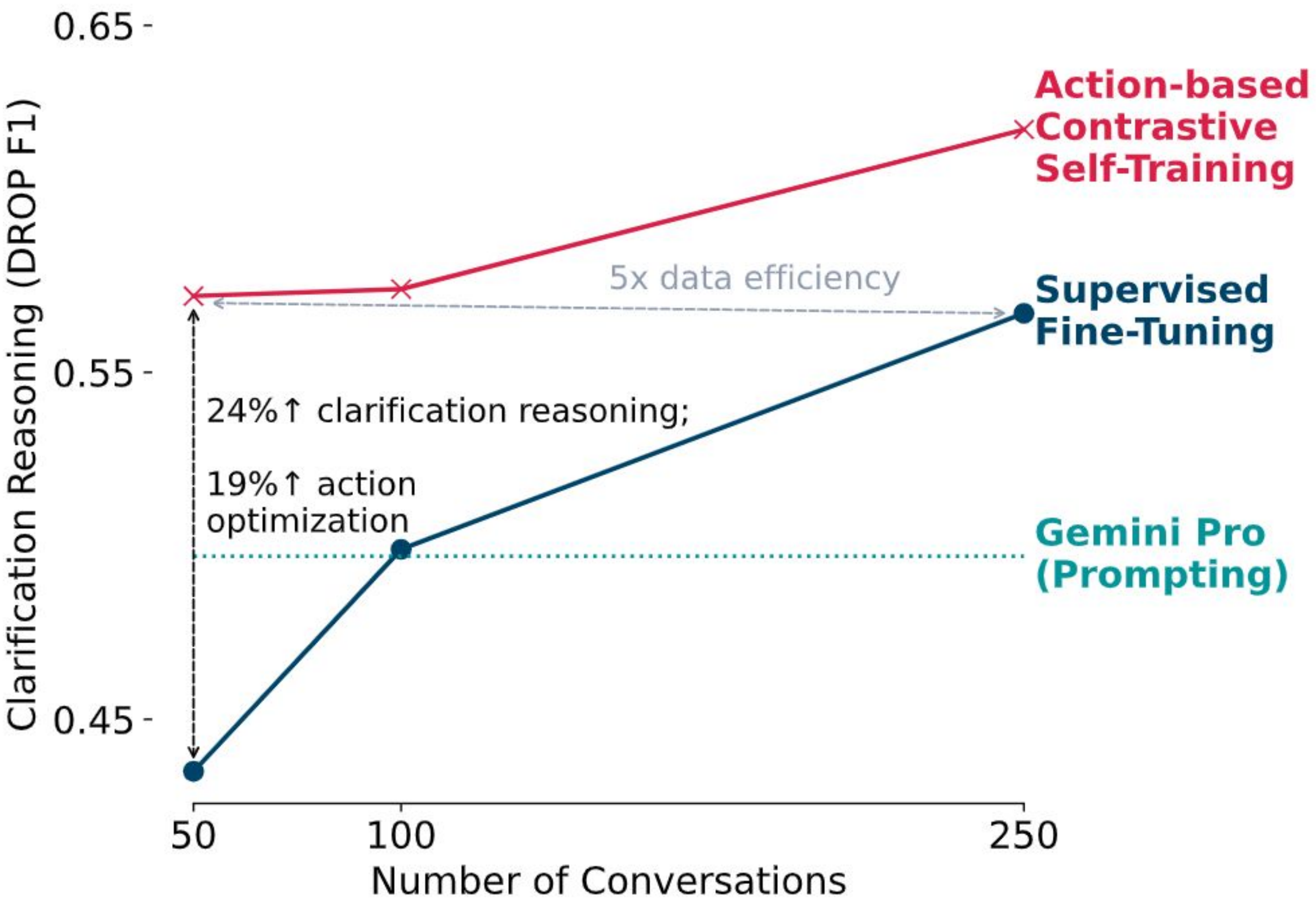
On-Policy Response Sampling



Prompt: "Show me information about airports related to the Robinson R-22 aircraft."
Winning Action: Clarify
Winning Response: What specific ...
Losing Response: `SELECT avg... FROM ...`
Losing Action: Answer
Information Goal: `SELECT avg...`

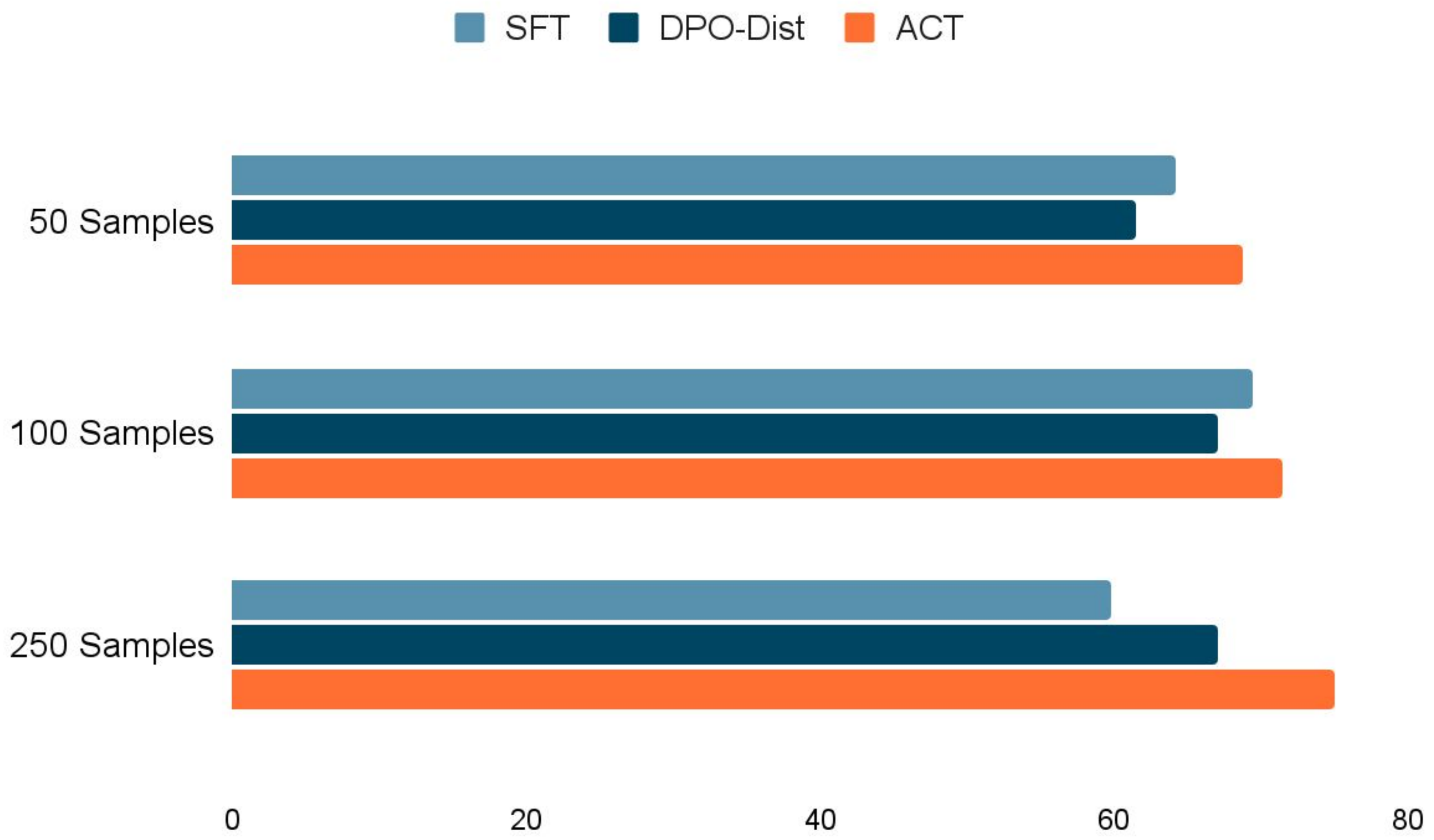
Policy at Step *i*

Action Classifier

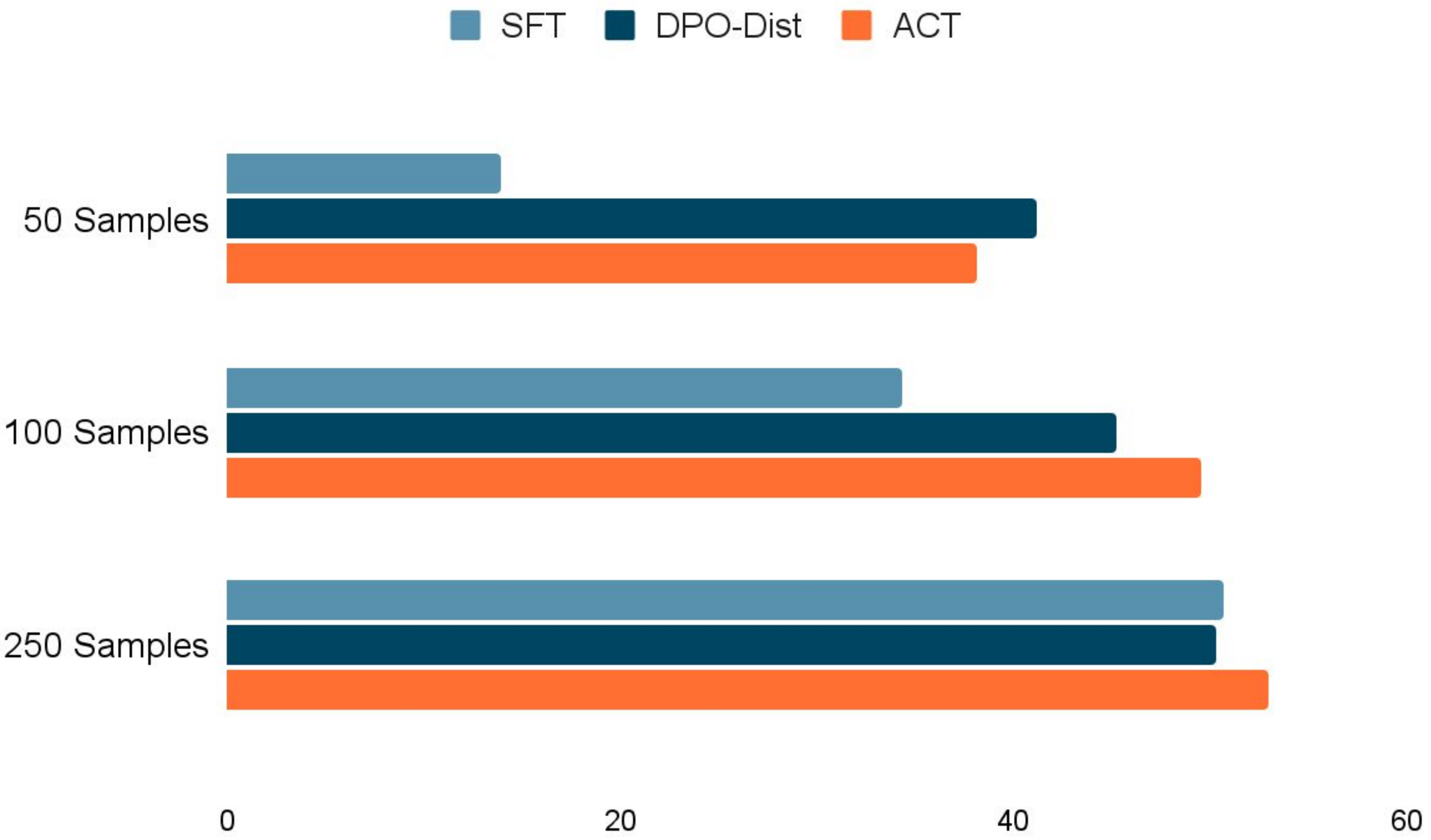


- Much more data efficient than standard tuning
- Outperforms frontier LLMs even with domain adaptation

Multiturn Similarity in Machine Reading Comprehension



Multiturn Execution Match in Ambiguous Text-to-SQL Generation



Takeaways

- Teaching conversational skills requires conversational objectives
- While SFT can do a lot with abundant data, RL approaches like ACT are more effective in the low data regime
- Building conversational models requires principled long-horizon optimization

Potential Future Directions

- Sophisticated execution feedback
- Large-scale ACT on unlabeled data
- Multimodality: spoken dialogue contains a lot of useful paralinguistic information

Questions?

Email: maxchen@cs.columbia.edu
Twitter: maximillianc_@

Questions?

Email: maxchen@cs.columbia.edu
Twitter: maximillianc_@