LLM-based Conversational Interaction

Mohammad Aliannejadi
University of Amsterdam

Conversational Interactions

- LLM-based conversational interactions improve query understanding:
  - Dynamic interactions,
  - Back-and-forth exchanges,
  - Clarify user intent,
  - Enhance search precision.
- Conversational search:
  - Unlike traditional search scenarios, builds context progressively,
  - Capture nuances,
  - Lead to more complex dynamics between the user and the system.

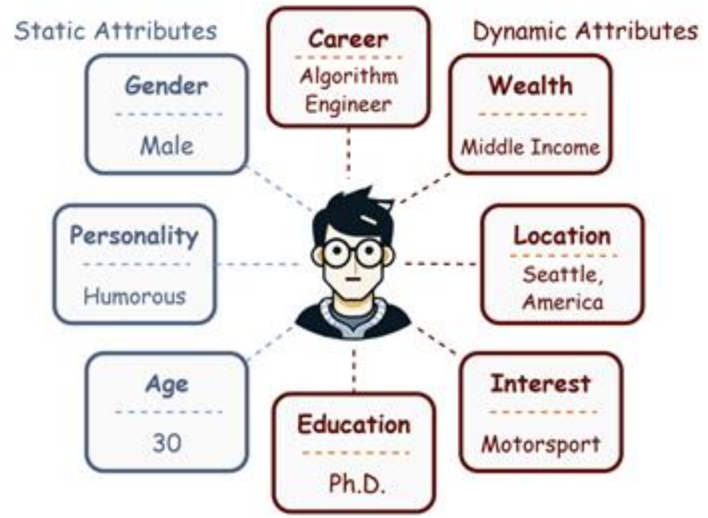
Interaction Simulation

- Complexity of user-system interactions.
- Scarcity of user data.
- Privacy concerns.
- More complex dynamics between user and system.
- Ability of the LLMs to perform multiple tasks and generalize well based on the massive training data.
- Use LLMs to simulate diverse behavior, intent, and query patterns.
- Help models learn to tackle complex queries and varied user needs effectively.

- See some example works that have tried this approach.

BASES: Web Search Simulation via LLM Agents

- User profile attributes:
  - Static
  - Dynamic
- Human and GPT-4 for user profile construction
  - Uniform attributes: gender
    - Random sampling
  - Non-uniform attributes: location
    - Sampling based on distribution
  - Unclear values: interest
    - Coarse-to-fine sampling



Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

BASES: Web Search Simulation via LLM Agents

- Agents simulate multi-turn search sessions:
  - Search
  - Click
  - Finish
- Preliminary results show that single prompt leads to bad results.
- Two prompting strategies:
  - Query Behavior Prompting
    - Keyword-based queries
    - Consider the information need and user profile
  - Click Behavior Prompting
    - Based on the top 10 results
    - Agents provide explanation why they decided to click on a page

Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

BASES: Results and Findings

- 90% consistency with real user query/click patter[illegible]
  - High realism
  - TREC-Session dataset
  - Compare query generation and rewriting behavior
  - Top-1 clicks of agent vs. real user
- Users with similar profiles behave differently.
  - Personalization preserved
  - Similar profiles with little differences tested on t[illegible] same information needs
  - Differences in the behavior was observed

What if the behavior variations were the result of re-generation variation or random seed?

Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

BASES: Results and Findings

- BASES-trained BERT outperforms real-user-trained models.
  - Model trained on simulated data vs. human-generated data.
  - Tested on English and Chinese benchmarks:
    - English data generated with GPT-4 and human annotation.
    - Chinese data sampled from Baidu search logs.

Methods	#Session (#Click)	*Chinese Benchmark*		
		MRR	**NDCG@1**	**NDCG@3**
BM25	-	45.16	27.20	41.39
BERT (TREC-Session)	1,257	-	-	-
BERT (AOL)	219,748	-	-	-
BERT (Tiangong-ST)	143,155	43.28	22.59	38.91
BERT (BASES)	1,000	<u>51.78</u>	<u>35.56</u>	<u>47.98</u>
BERT (BASES)	10,000	**53.52**	**35.98**	**50.72**

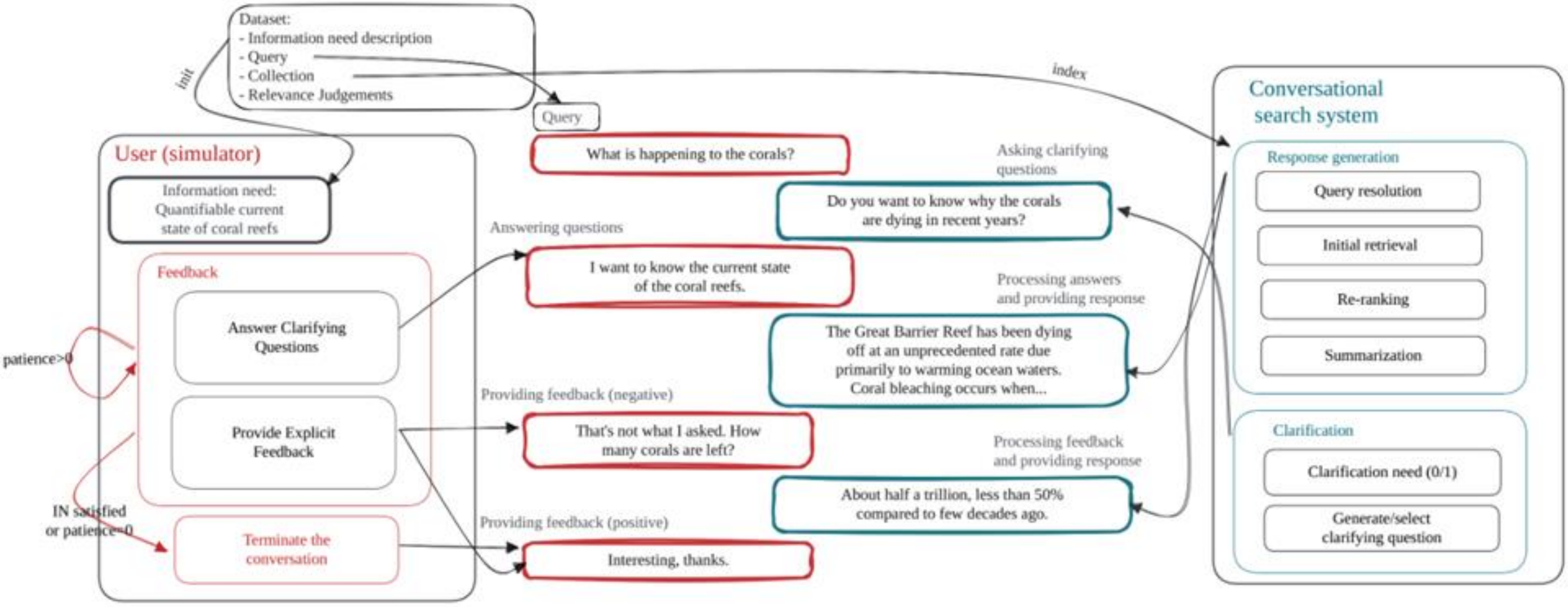
Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

Simulation of Mixed-initiative Interactions

- LLM-based user simulation for mixed-initiative scenarios:
  - USi: based on GPT-2
  - ConvSim: based on GPT-3
- More diverse set of simulated user actions:
  - Provide explicit feedback
  - Answer clarifying questions
  - Engage in a multi-turn information-seeking conversation

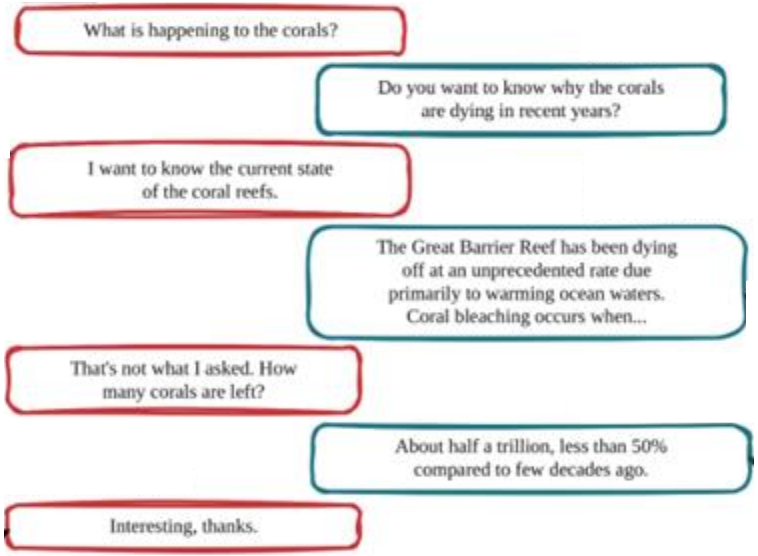
Sekulić et al., "Evaluating Mixed-initiative Conversational Search Systems via User Simulation", (WSDM '22)
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



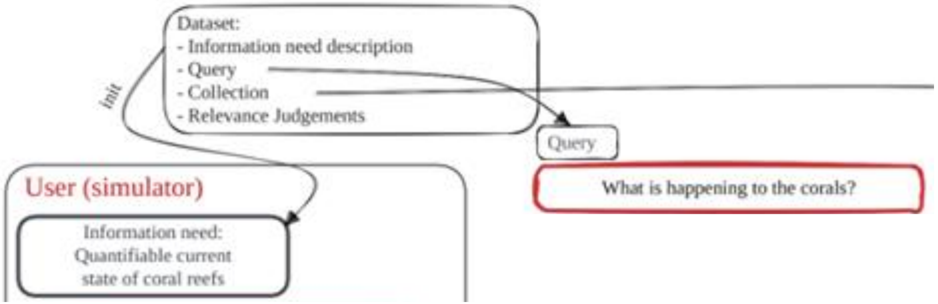
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



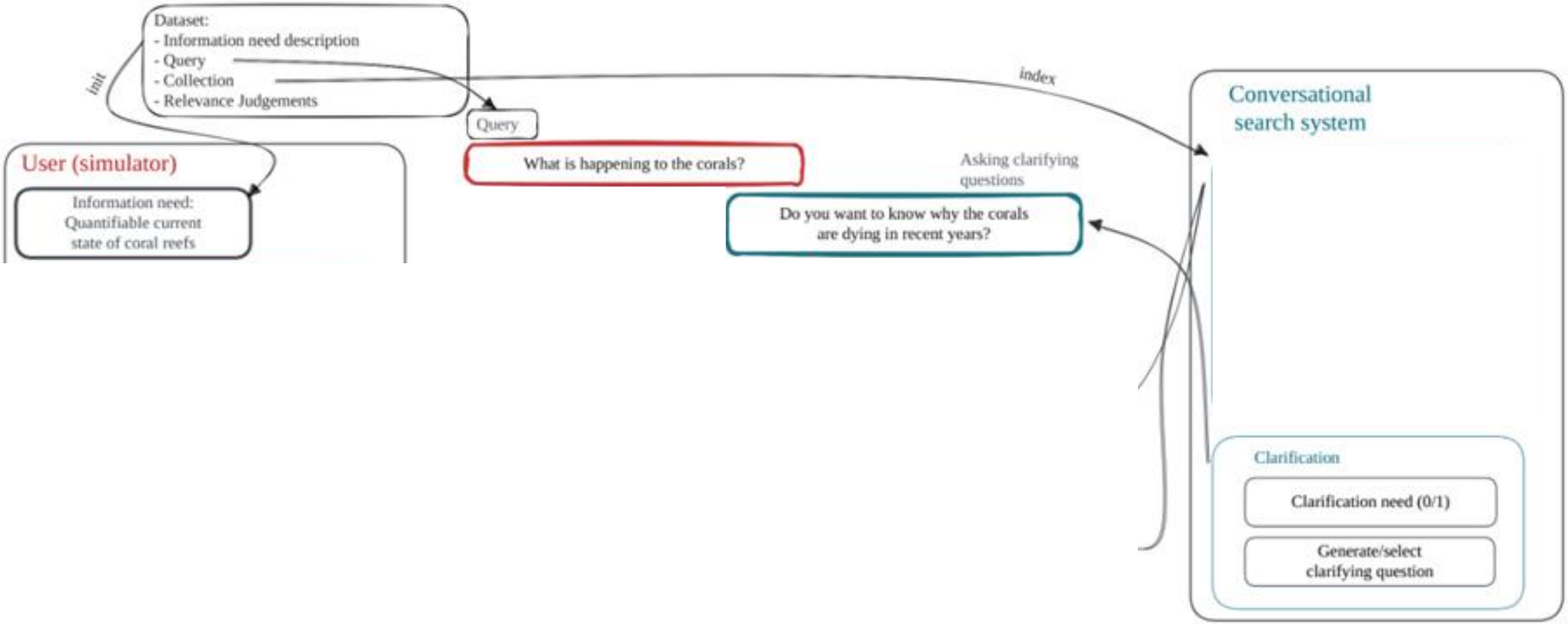
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



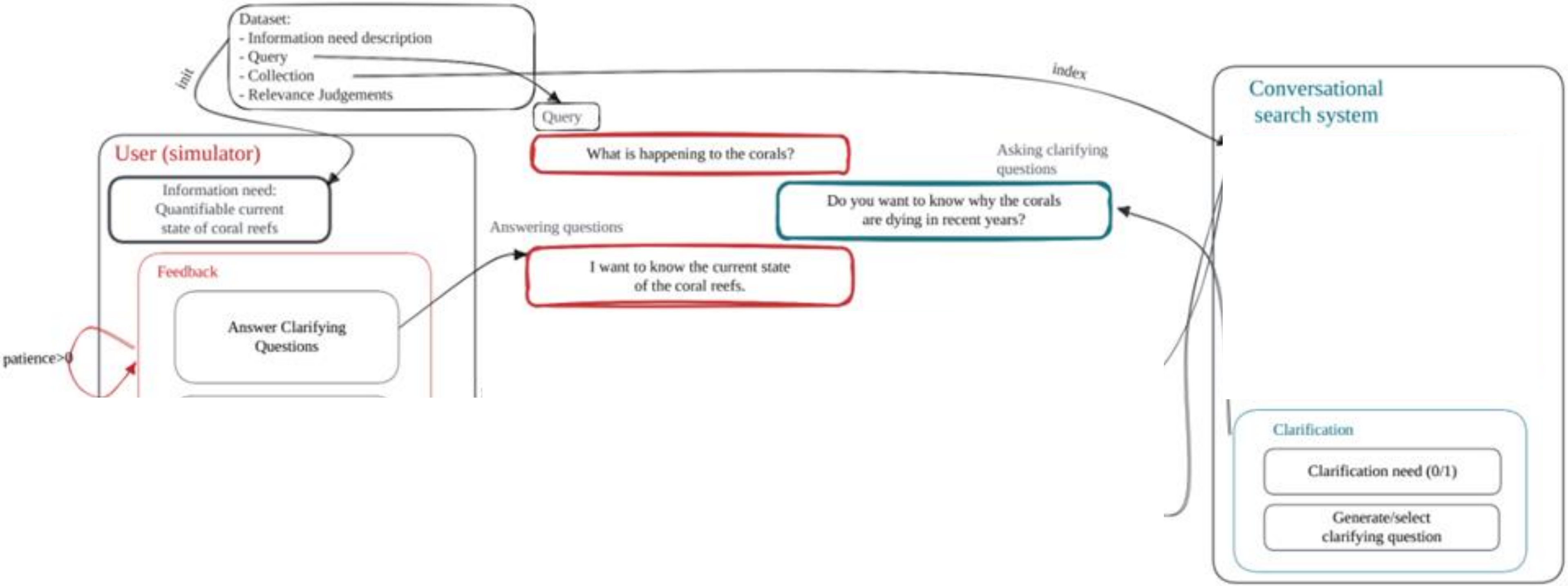
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



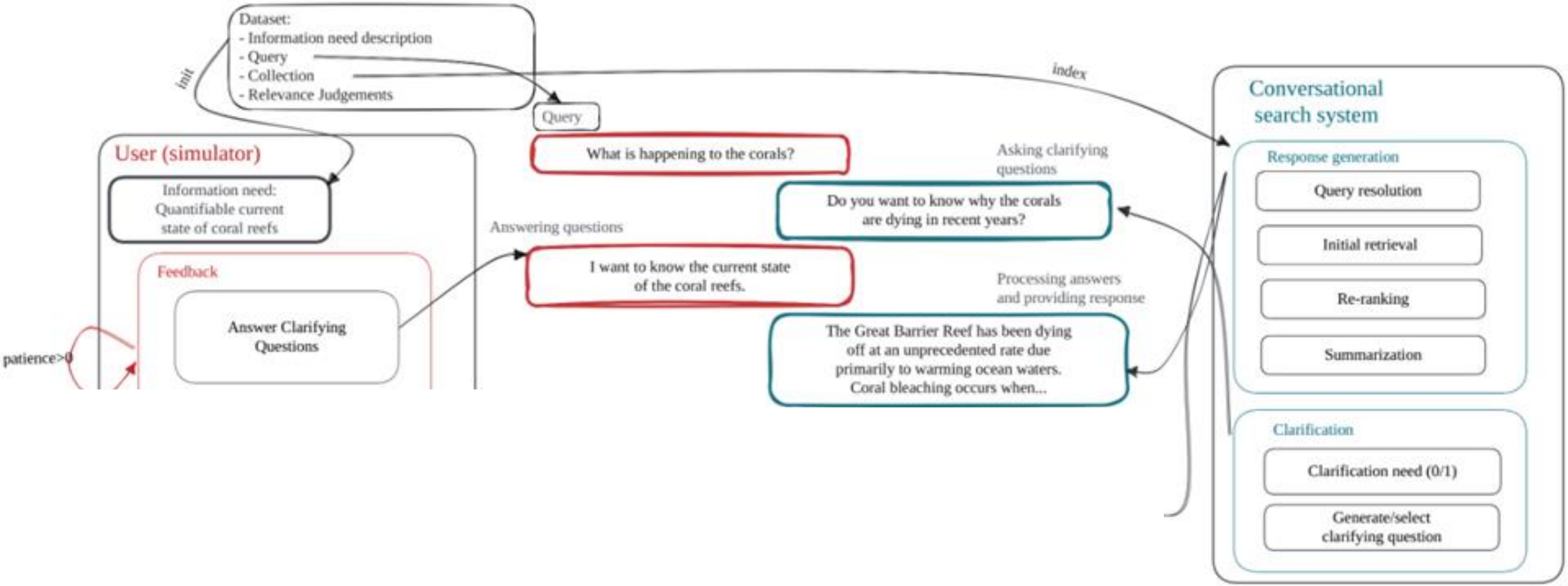
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



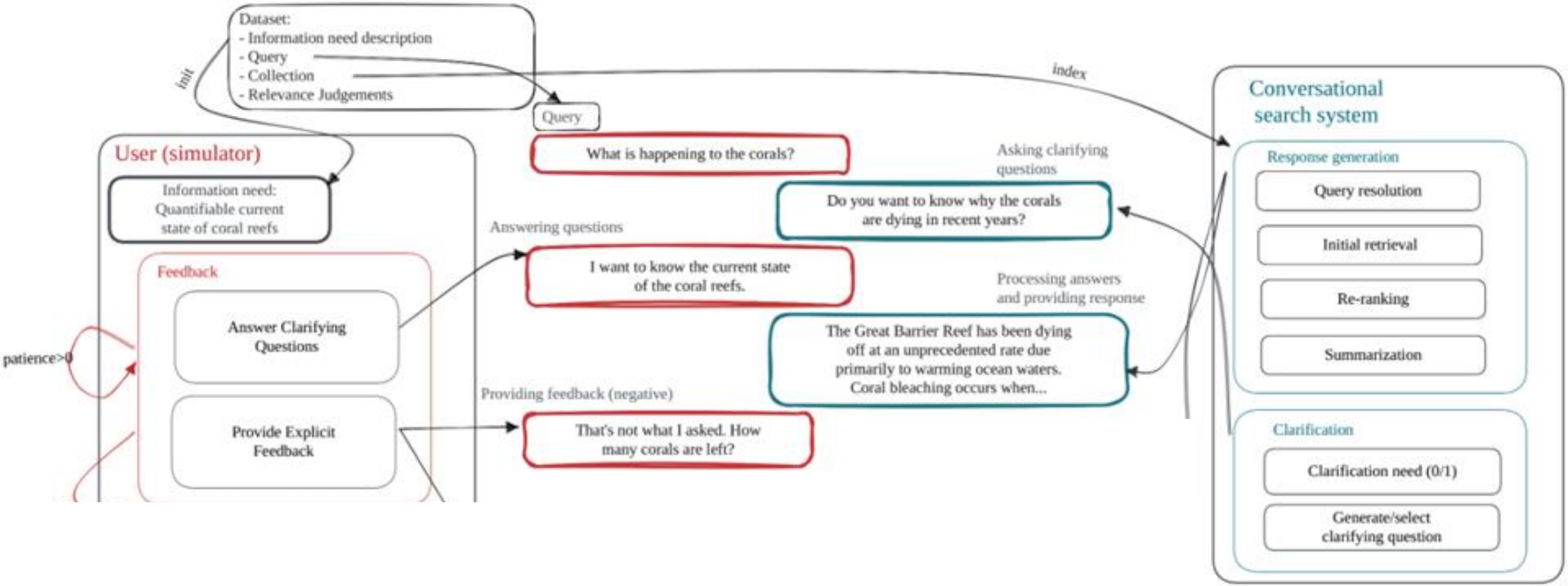
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



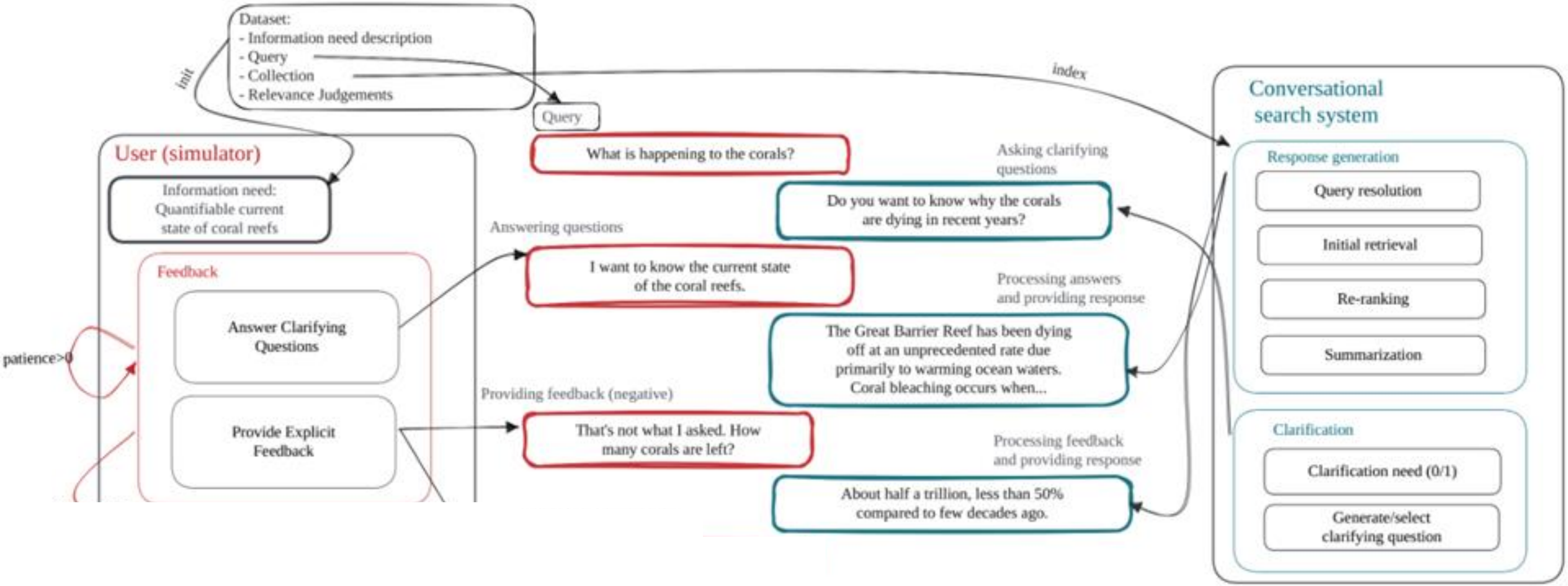
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



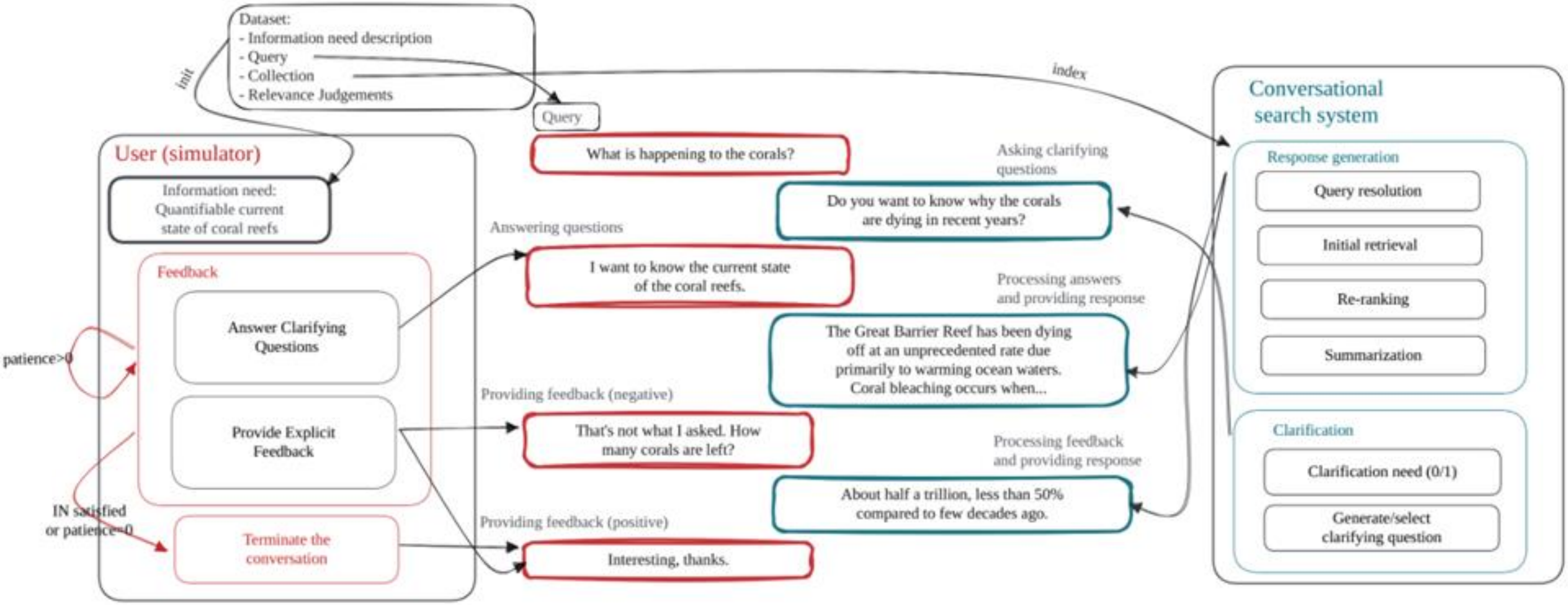
Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim



Owoicho et al., "Exploiting Simulated User Feedback for Conversational Search: Ranking, Rewriting, and Beyond", (SIGIR '23)

ConvSim: Performance

		ConvSim [44]	*USi* [61]	Ties	ConvSim [44]	Human	Ties
Single	*Naturalness*	37%†	22%	41%	36%	25%	39%
	Usefulness	44%†	19%	37%	36%†	20%	44%
Multi	*Naturalness*	45%†	18%	37%	25%	28%	47%
	Usefulness	62%†	12%	26%	26%	16%	58%

Sekulić et al., "Analysing Utterances in LLM-Based User Simulation for Conversational Search", (ACM TOIS'24)

ConvSim: Performance

	Human	*USi* [61]	*ConvSim* [44]
Repeat	2%	0%	3%
Repeat/rephrase	4%	7%	6%
Repeat/simplify	4%	8%	5%
Clarify/refine	63%	37%	83%
Other	25%	40%	3%
Hallucination	2%	7%	0%

Sekulić et al., "Analysing Utterances in LLM-Based User Simulation for Conversational Search", (ACM TOIS'24)

Going Smaller with Finetuning

- Smaller finetuned LLMs can be better in simulation

- Task-oriented dialogues
- Domain awareness
- Finetuning Llama-2 13B

- Less hallucination
- Improved user intent alignment

Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)
Sekulić et al., "Reliable LLM-based User Simulator for Task-Oriented Dialogue Systems", (SCI-CHAT 2024)

Finetuning for Feedback

- Finetuning T5 leads to a strong baseline

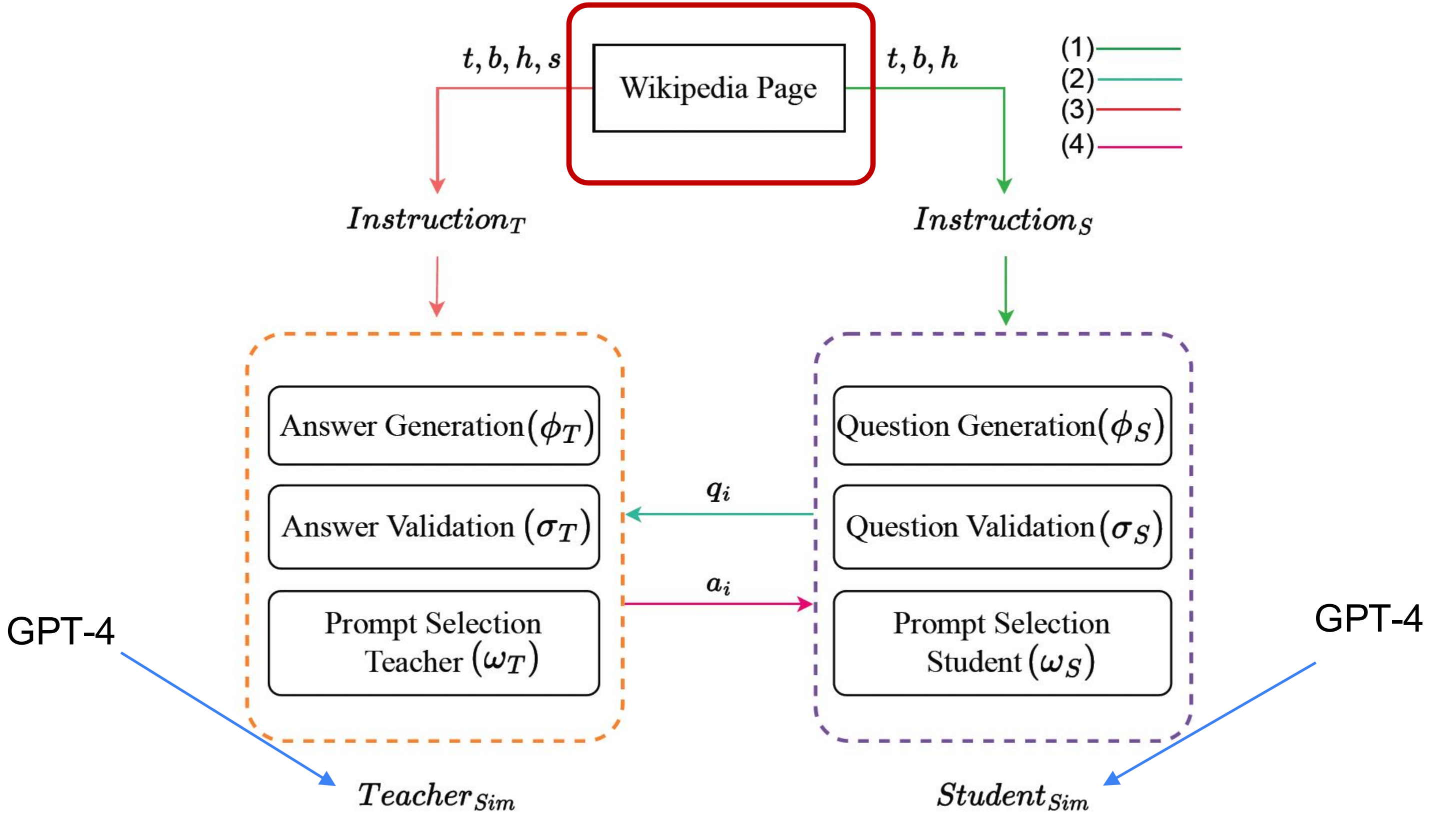
Dataset	Model	Generation Metrics			
		BLEU-3	BLEU-4	ROUGE-L	METEOR
Qulac	GPT-2 (USi[47])	12.6	9.1	28.2	28.9
	GPT-3.5 (ConvSim [33])	13.5	9.8	29.1	29.0
	T5-small	**23.7**†	**19.0**†	**40.8**†	**43.2**†
ClariQ	GPT-2 (USi[47])	13.5	9.8	28.8	28.6
	GPT-3.5 (ConvSim [33])	13.4	9.7	28.9	28.4
	T5-small	**24.3**†	**19.5**†	**41.0**†	**43.3**†

Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

Proactive Simulation

- User simulator reacts to the system's action
- An actual user would be more proactive:
  - Start a conversation
  - Steer the topic exploration by asking further questions
- Leverage LLMs to explore a given topic, as well as reacting to the system's response
- QuAC-like setup:
  - Student: knows very little about a topic and aims to learn more about it
  - Teacher: knows much more and provides response to the Student, based on a provided document
- Replace crowd workers of QuAC with LLMs

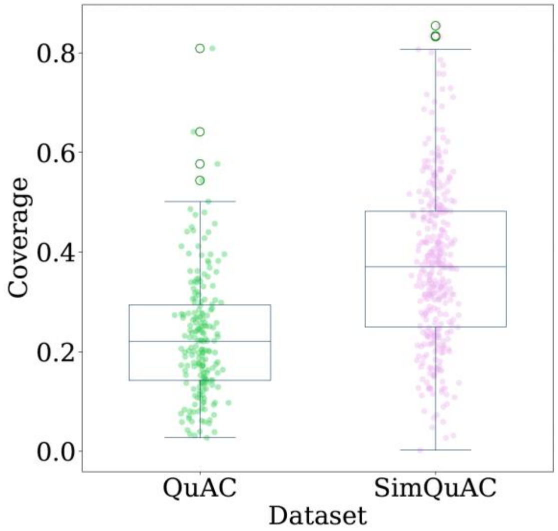
Abbasiantaeb et al., "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions", (WSDM'24)
Choi et al., "QuAC: Question Answering in Context", (EMNLP '18)



Abbasiantaeb et al., "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions", (WSDM'24)

SimQuAC

- Simulated conversational question-answering dataset
- Compare with crowd-sourced dialogues of QuAC
  - More natural dialogue flow
  - More effective exploratory behavior: more subtopics are covered



Limitations of Simulation

- In-depth analysis of where simulation fails
- Misalignment
- Errors
- Noise
- Evaluation limitations
- Case study on Qulac and ClariQ

Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

Limitations of Simulation

- In-depth analysis of where simulation fails
  - Different types of error discussed

i = "Find the homepage of the president of the United States."

i = "How do I register to take the SAT exam?"
q = "SAT"

i = "I'm looking for websites that do antique appraisals"
q = "Appraisals"

i = "Find information on various types of computer memory, and how they are different."
q = "Memory"
cq = "Who was the first to study the brain and memory?"
G = "I want to know how different they are."
H = "Herman Ebbinguaus."

Reasons	T5
Wrong answer type	33.9%
Cooperativeness mismatch	31.1%
Both valid	13.9%
Extra information	10.3%
Noisy reference	5.8%
Miscellaneous	4.2%
Total # ROUGE<0.2	360

Wang et al., "An In-depth Investigation of User Response Simulation for Conversational Search", (WWW '24)

Creating Interaction Sandbox

- LLM-based agent
- Sandbox environment
- High similarity to human behavior
- Study two social phenomena:
  - (i) information cocoons
  - (ii) user conformity behaviors.

Profile Module

ID	Name	Age	Gender	Career	Traits	Interest
1	David	25	Male	Doctor	Caring	Action

Wang et al., “User Behavior Simulation with Large Language Model-based Agents”, ACM TOIS, 2025

Information Cocoon

- Users only access information similar to their own preference, but los the opportunity to view more diverse options.
- Information cocoon is measured by entropy:

# of categories for u

$$E = -\frac{1}{|U|} \sum_{\boxed{u \in U}} \sum_{\boxed{c \in C}} \boxed{f_{u,c}} \log f_{u,c}$$

User set

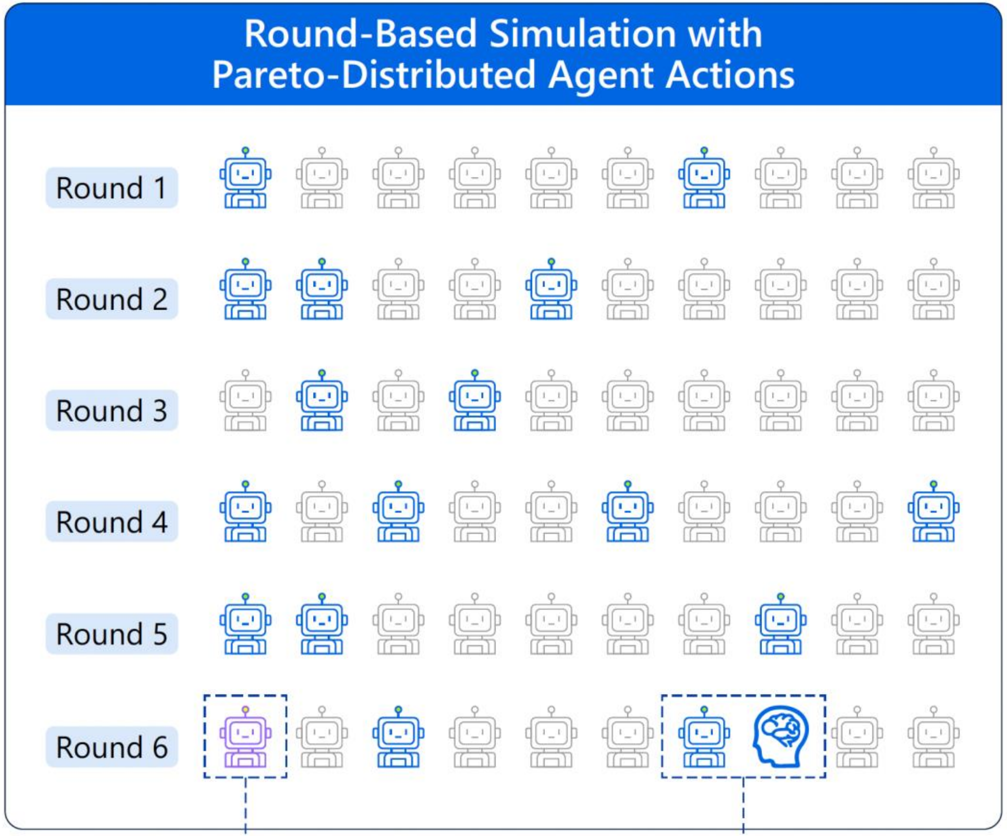
Category set

- Smaller E indicates more severe information cocoon

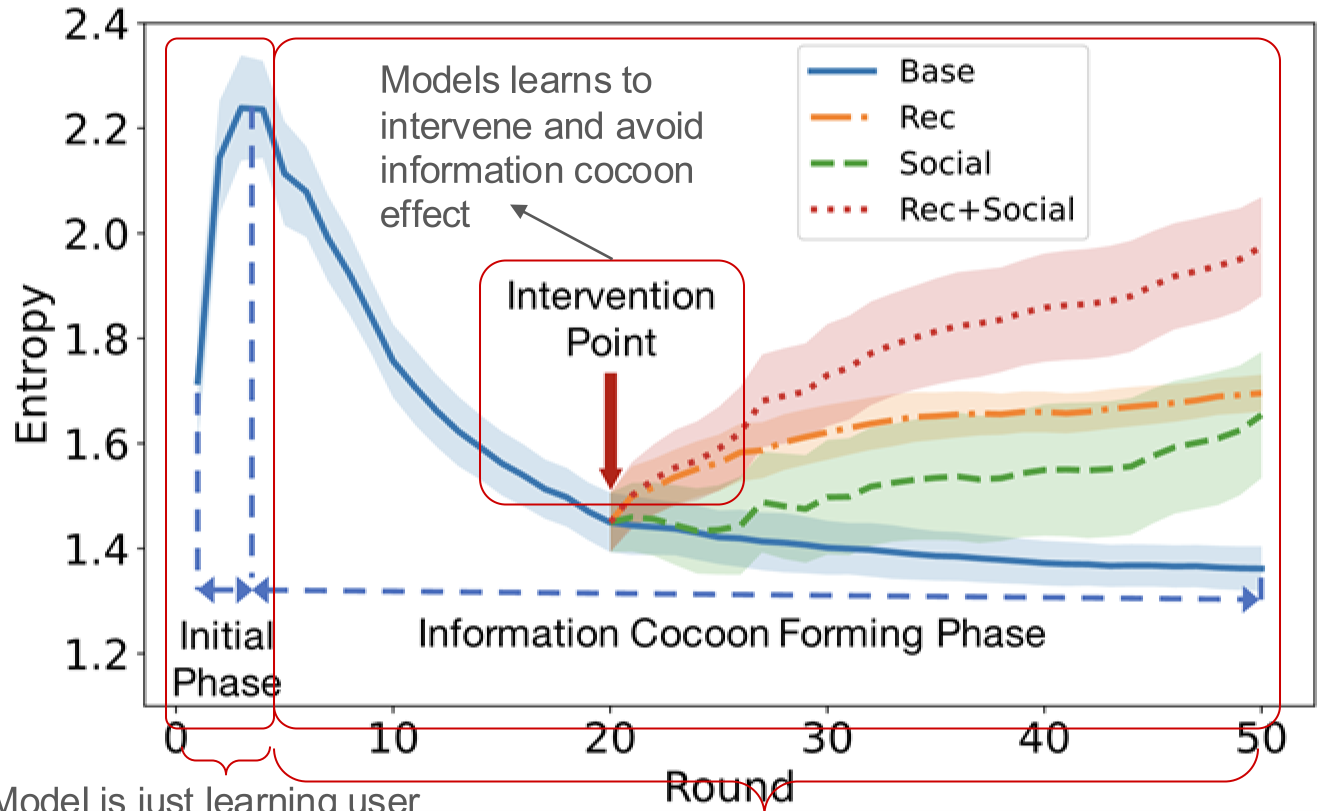
Nguyen et al., "Exploring the filter bubble: the effect of using recommender systems on content diversity", WWW '14
Paio et al., "Human–AI adaptive dynamics drives the emergence of information cocoons", Nature Machine Intelligence, 2023

Simulating Information Cocoons

- Simulate information cocoon and propose solutions for it.
- Recommender systems:
  - 50 agents freely interact with items, leading to an agent-item matrix.
  - Recommender systems trained at each round of interaction for 50 times.
  - Information cocoon is measured by entropy.



Wang et al., "User Behavior Simulation with Large Language Model-based Agents", ACM TOIS, 2025



Model is just learning user preference and recommends diverse items

Model start learning user preferences and recommendations become less diverse

Wang et al., "User Behavior Simulation with Large Language Model-based Agents", ACM TOIS, 2025

Implications

- How can LLM-based simulation be used to improve query understanding?
- LLMs to be used to generate query variations – literature has shown the potential
- Simulation can be informed by the current human variation studies
- Create sandbox environment
- Enrich existing datasets
- Incorporate query variations in the existing simulators

Alaofi et al., "Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study", SIGIR '23
Abu Onq et al., "Classifying Term Variants in Query Formulation", SIGIR '25

Questions