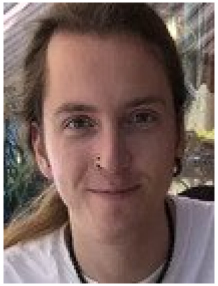
Simulating Follow-up Questions in Conversational Search

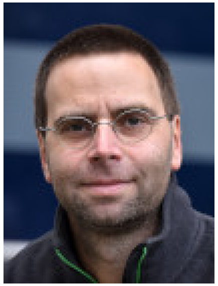
ECIR 2024

Johannes Kiesel

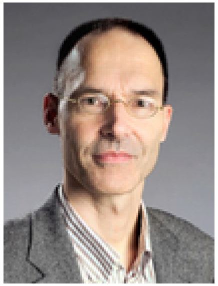


Marcel Gohsen

Nailia Mirzakhmedova



Matthias Hagen



Benno Stein

Bauhaus-Universität Weimar

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Webis

Conversational Search Systems

What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

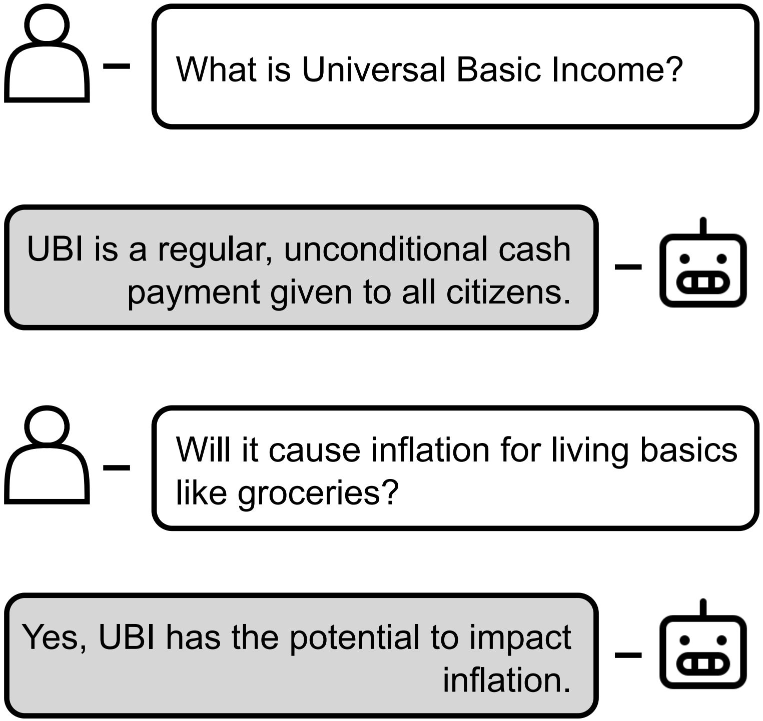
Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Conversational Search Systems

Evaluation

How to evaluate conversational search systems?



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

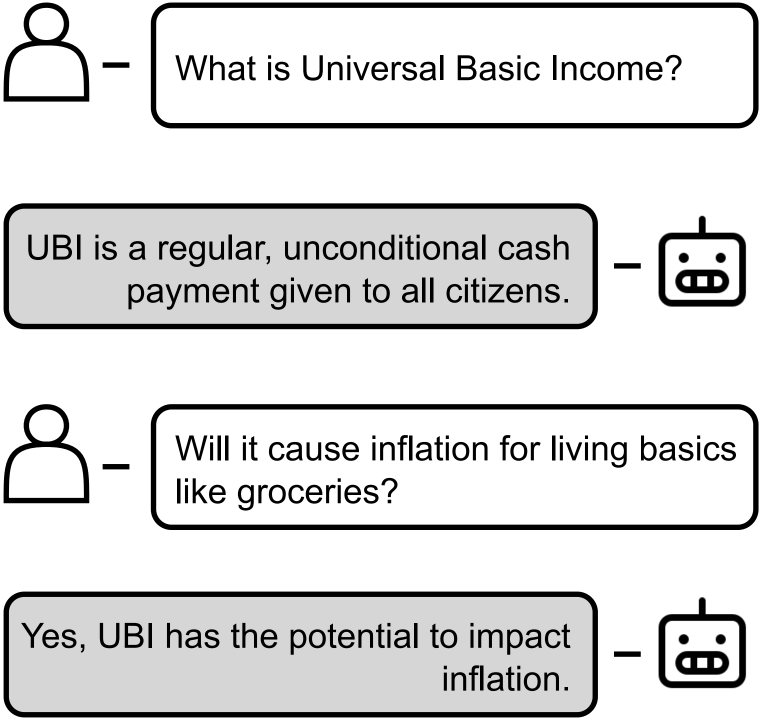
Yes, UBI has the potential to impact inflation.

Conversational Search Systems

Evaluation: User Studies

Problems with real user evaluation:

1. Expensive
2. Time-consuming
3. Requires a live service
4. Quality
5. Scalability
6. Reproducibility
7. Ethical concerns



What is Universal Basic Income?

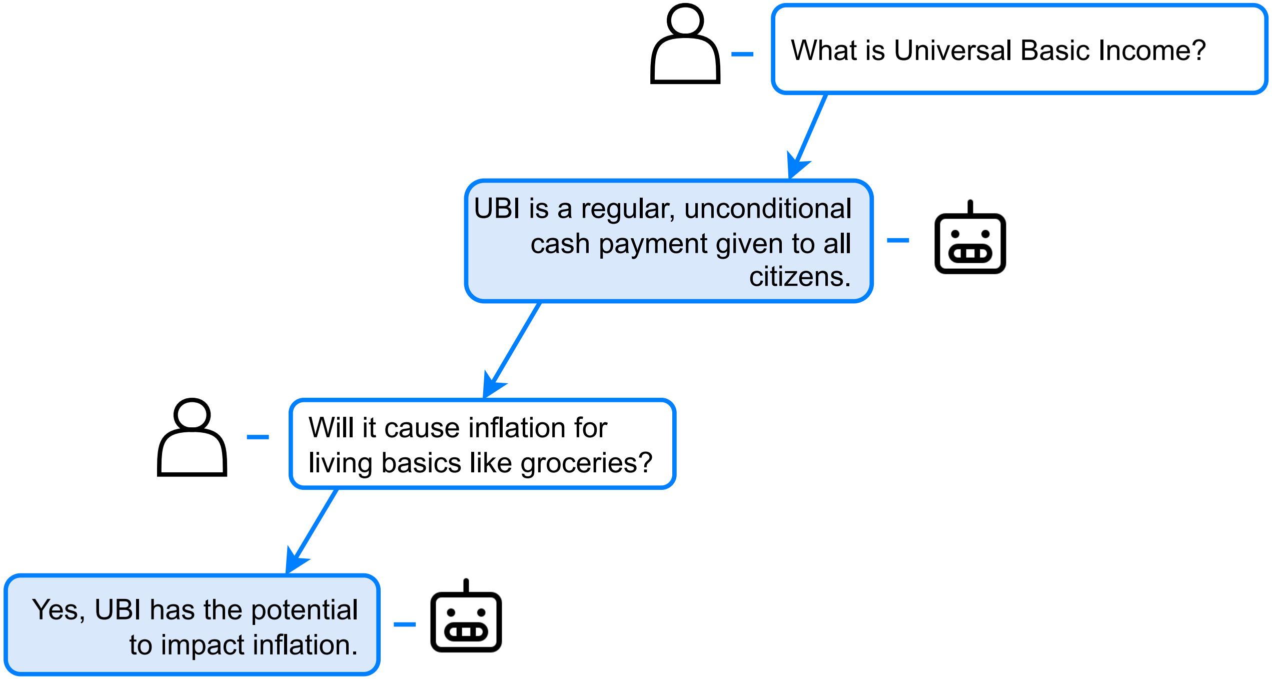
UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Conversational Search Systems

Evaluation: Reusable Test Collections



What is Universal Basic Income?

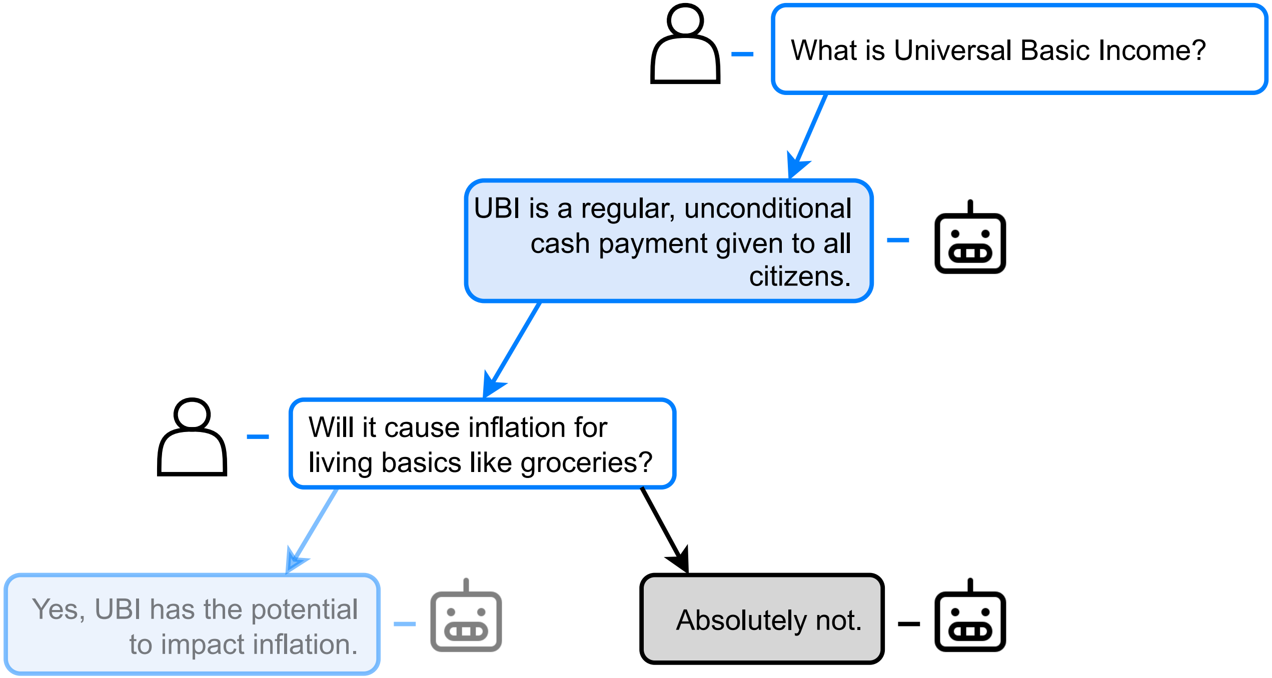
UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Conversational Search Systems

Evaluation: Reusable Test Collections



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

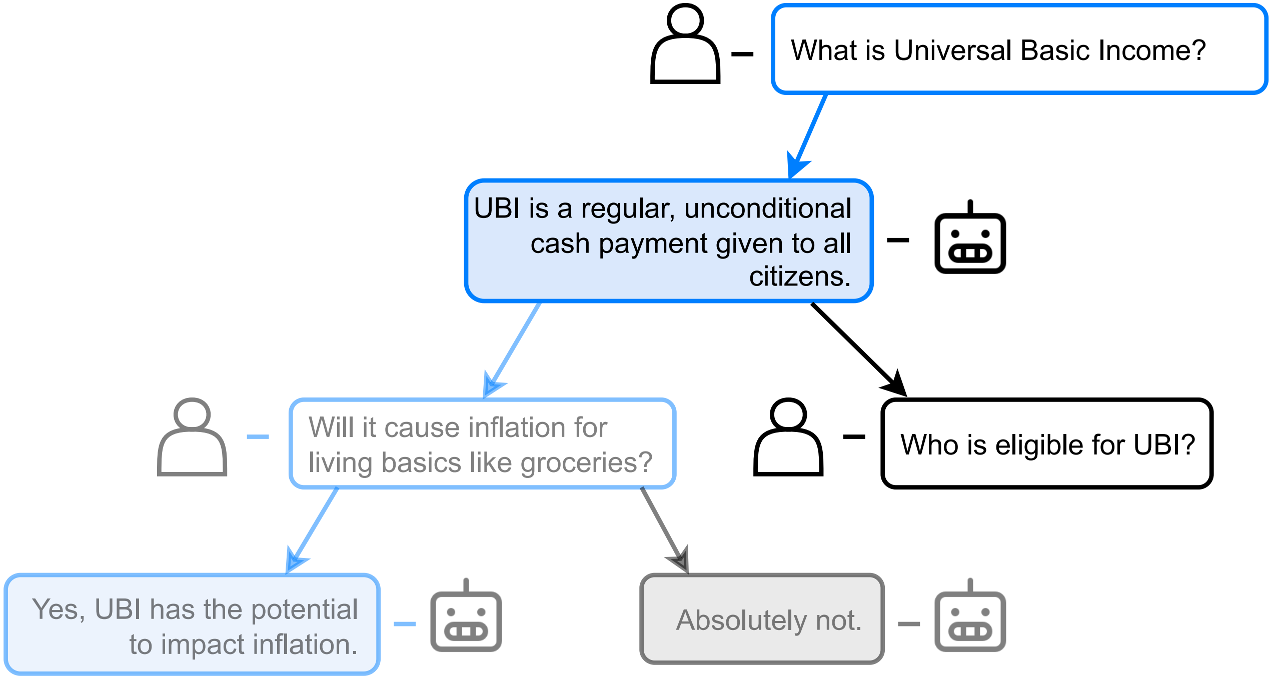
Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Absolutely not.

Conversational Search Systems

Evaluation: Reusable Test Collections



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

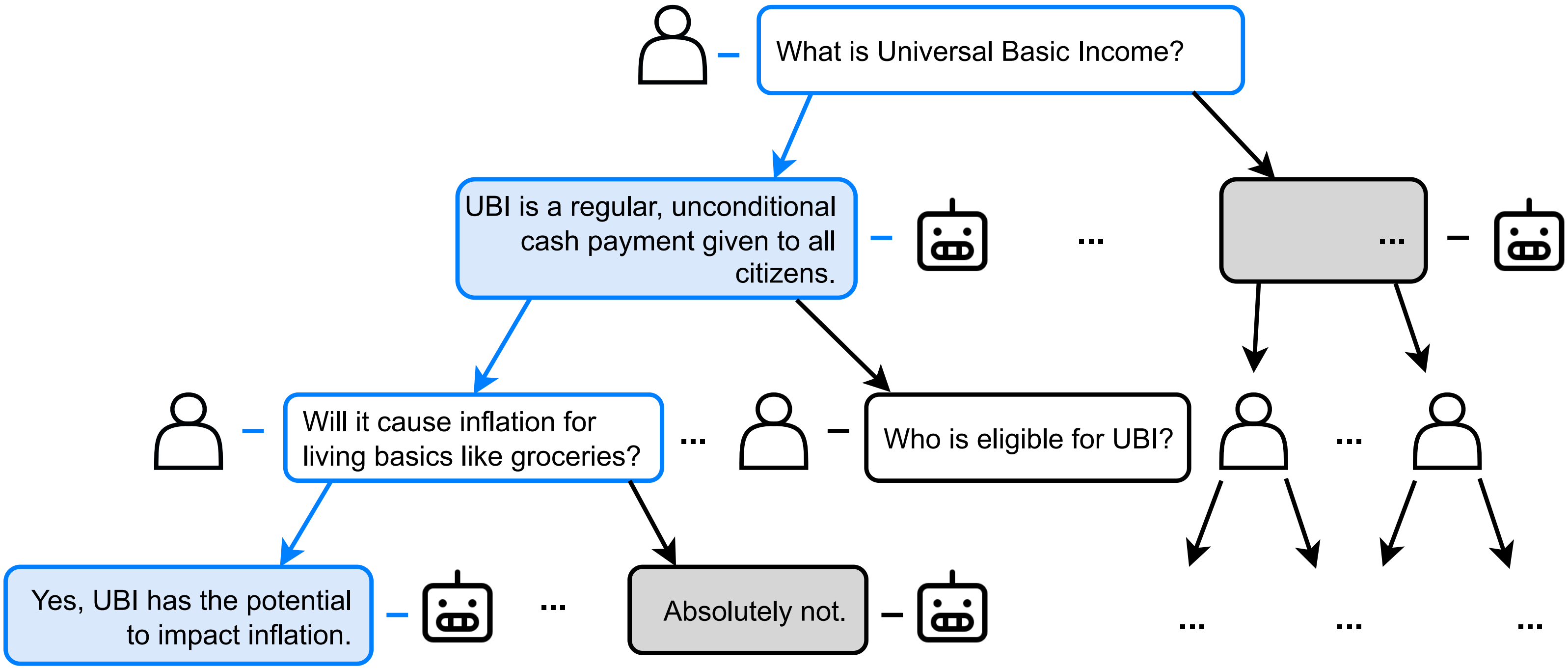
Who is eligible for UBI?

Yes, UBI has the potential to impact inflation.

Absolutely not.

Conversational Search Systems

Evaluation: Reusable Test Collections


What is Universal Basic Income?
UBI is a regular, unconditional cash payment given to all citizens.
...
...
Will it cause inflation for living basics like groceries?
...
Who is eligible for UBI?
...
Yes, UBI has the potential to impact inflation.
...
Absolutely not.
...
...
...

Problems with static test collection evaluation:

1. The system is limited in selecting the best response
2. The space of possible dialogue states increases exponentially

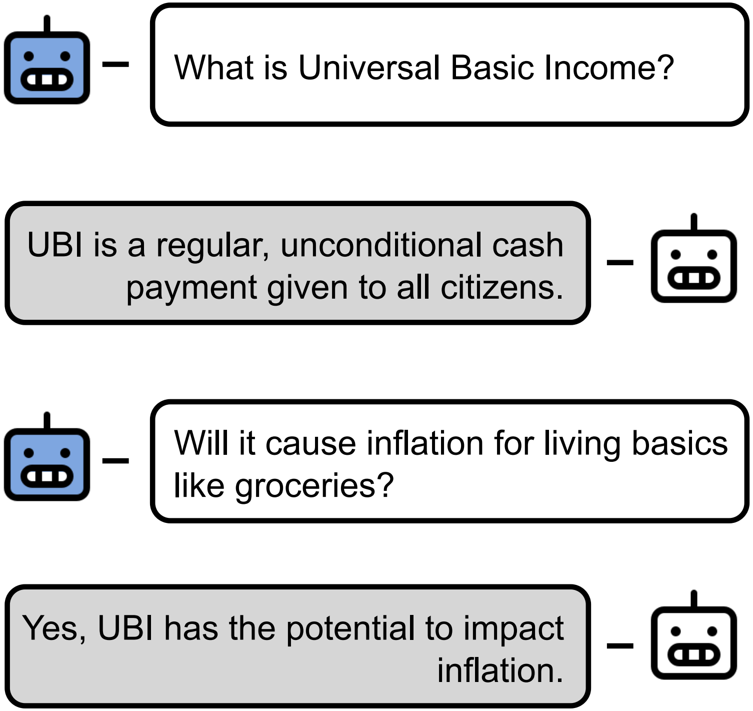
Conversational Search Systems

Evaluation: User Simulation

Idea: Replace 👤 with 🤖
(including judgement)

Advantages:

1. Scalability
2. Reproducibility
3. Cost effective
4. Dynamic
5. Control over scenarios



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

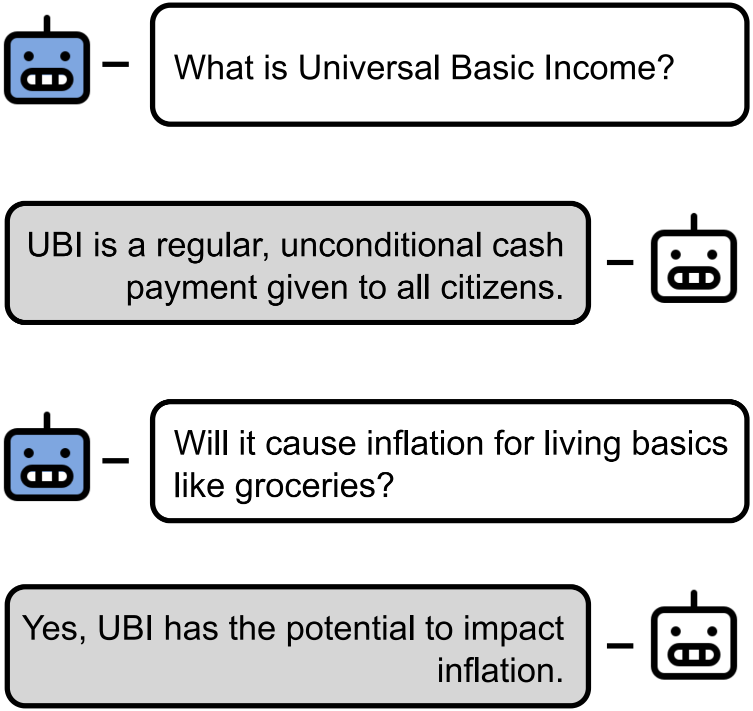
Conversational Search Systems

Evaluation: User Simulation

Idea: Replace 👤 with 🤖
(including judgement)

Advantages:

1. Scalability
2. Reproducibility
3. Cost effective
4. Dynamic
5. Control over scenarios



🤖 – What is Universal Basic Income?

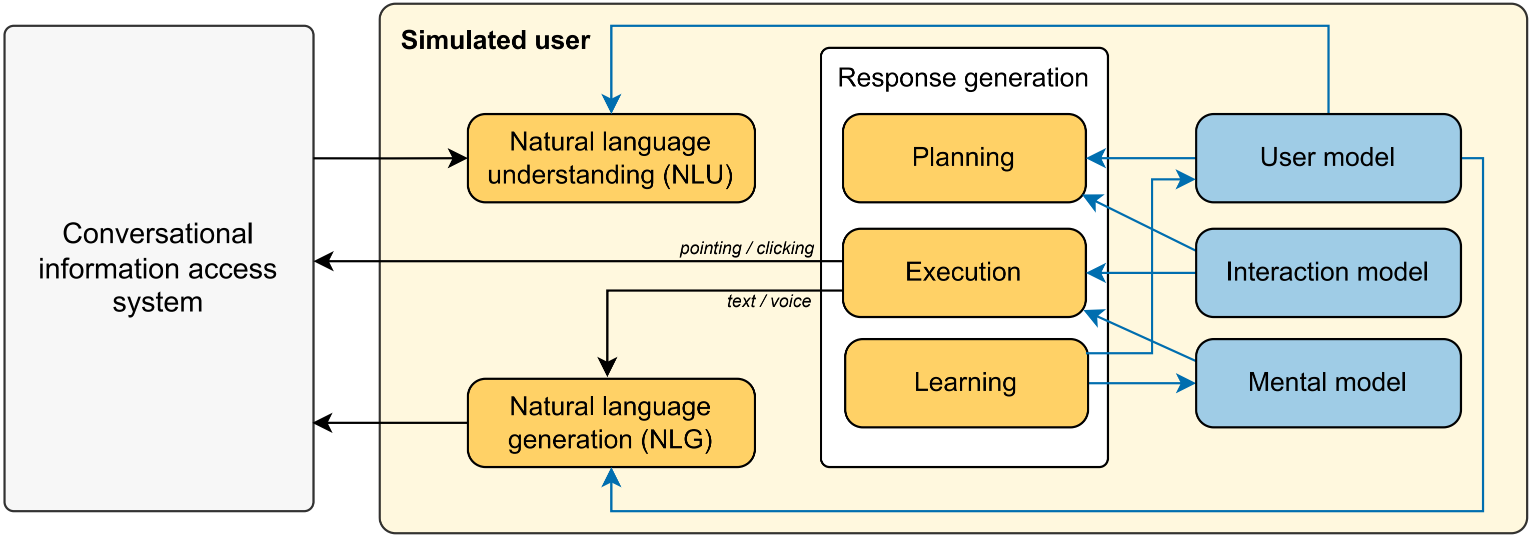
UBI is a regular, unconditional cash payment given to all citizens. – 🤖

🤖 – Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation. – 🤖

Conversational Search Systems

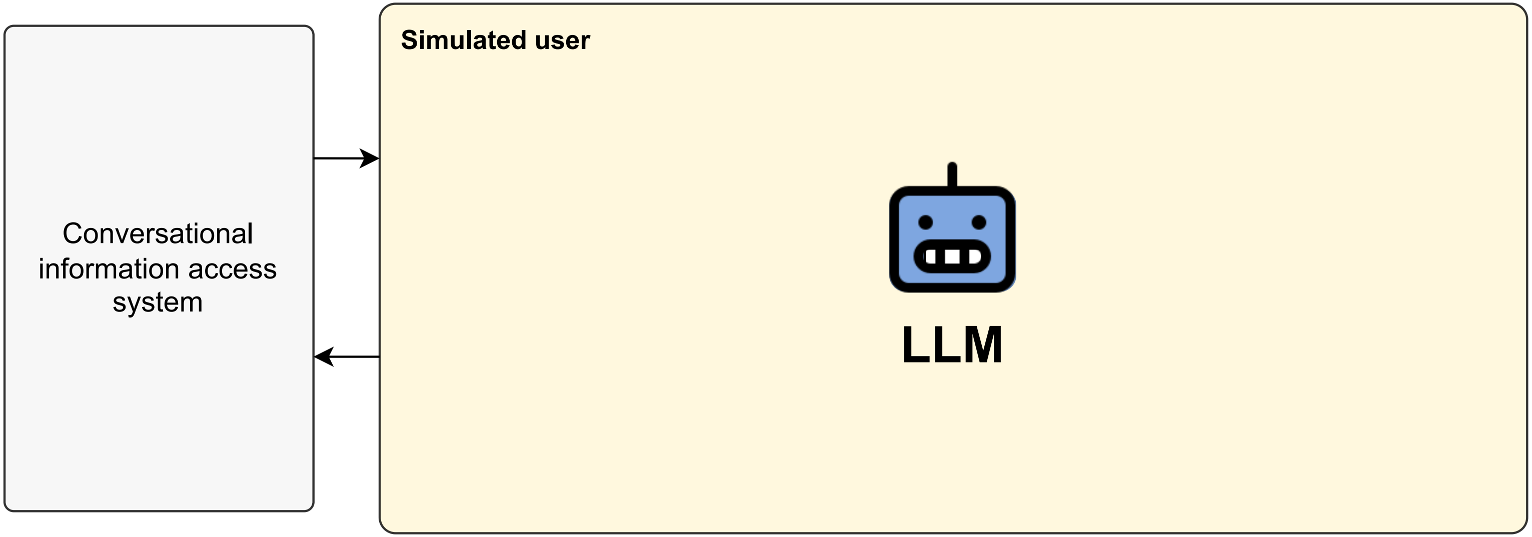
User Simulation Model



Conceptual architecture of a user simulator proposed by K. Balog (2021)

Conversational Search Systems

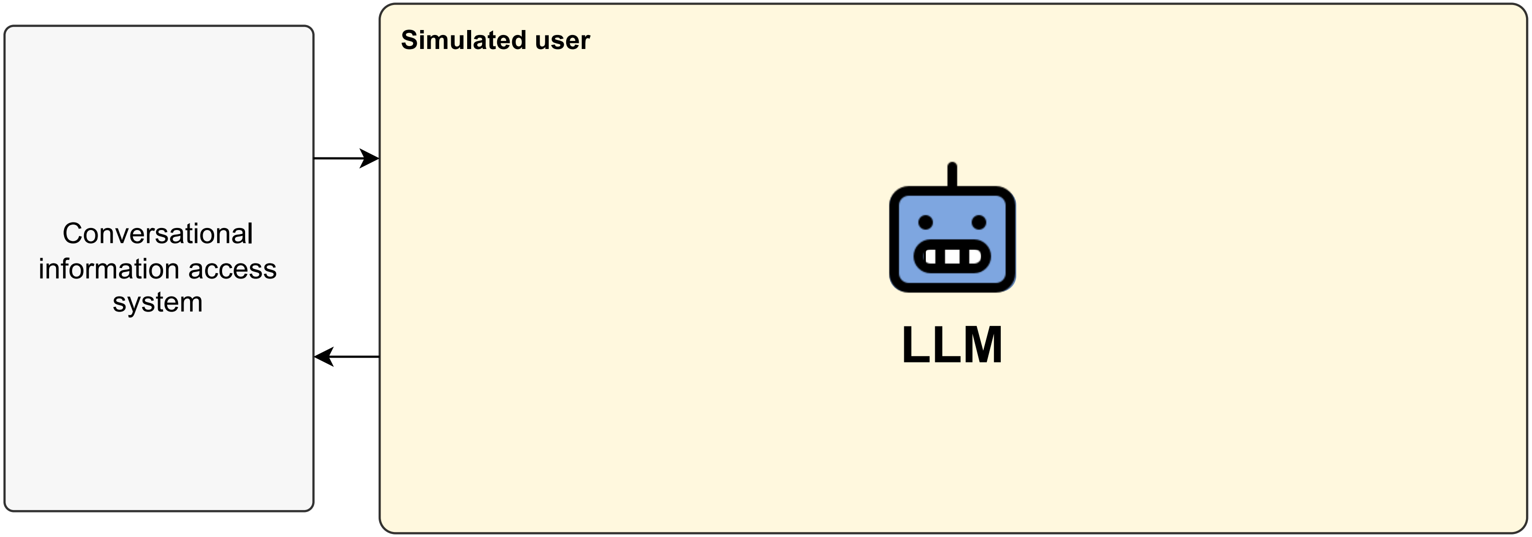
User Simulation Model



If an LLM can simulate someone that you chat with, can it also simulate someone that searches?

Conversational Search Systems

User Simulation Model



If an LLM can simulate someone that you chat with, can it also simulate someone that searches?

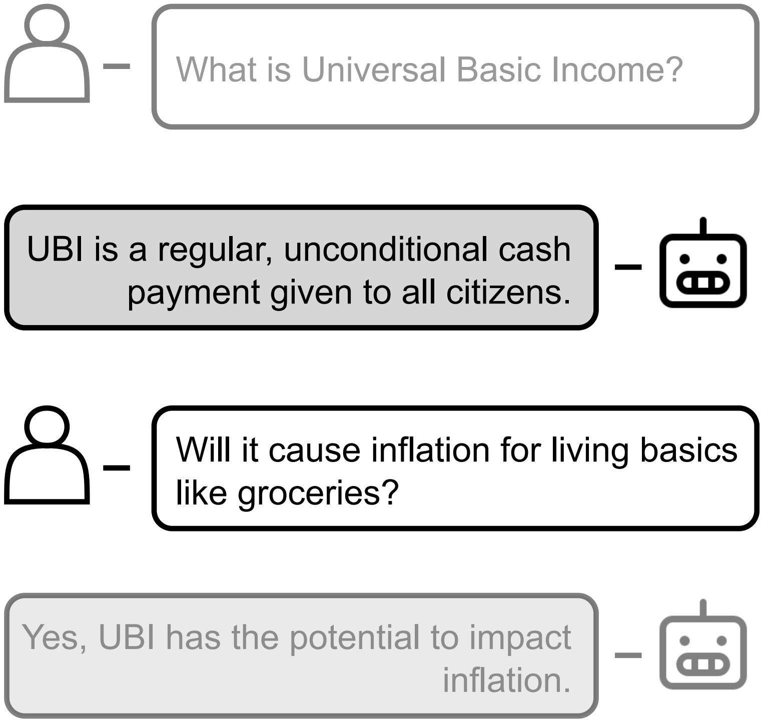
Simulating Follow-up Questions in Conversational Search

Follow-up Questions

Follow-up questions:
Questions about something the search system said earlier.

❑ Conversational QA

❑ Conversational Search
e.g. around 54% of user utterances in TREC CAsT 2022 [Owoicho et al., 2022] are follow-up questions



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

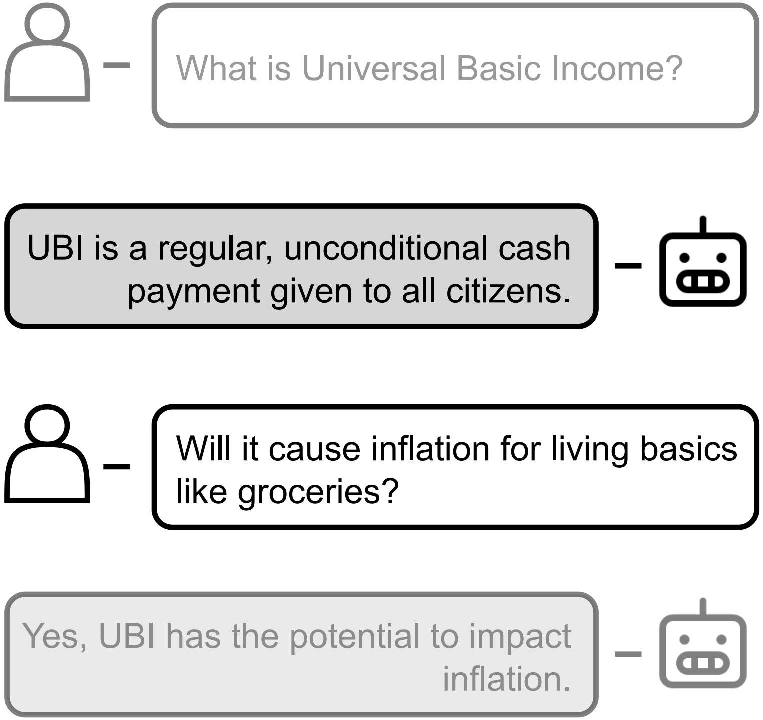
Yes, UBI has the potential to impact inflation.

Simulating Follow-up Questions in Conversational Search

Follow-up Questions

Follow-up questions:
Questions about something the search system said earlier.

- □ Conversational QA
- □ Conversational Search
  e.g. around 54% of user utterances in TREC CAsT 2022 [Owoicho et al., 2022] are follow-up questions



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Can LLMs simulate user follow-up questions?

Simulating Follow-up Questions in Conversational Search

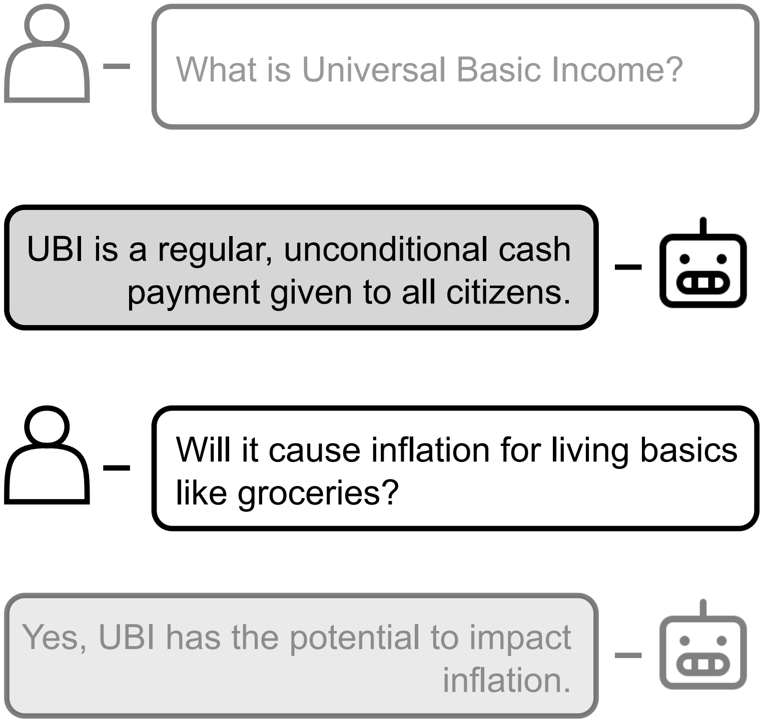
Follow-up Questions

Task: Given an informative textual response to a user's query, generate a question a user might ask based on the provided information.

RQ1. Are the generated questions similar to human questions?

RQ2. According to human judgments, are the generated questions appropriate follow-up questions?

RQ3. Can simple prompt modification result in simulation of different user profiles?



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Simulating Follow-up Questions in Conversational Search

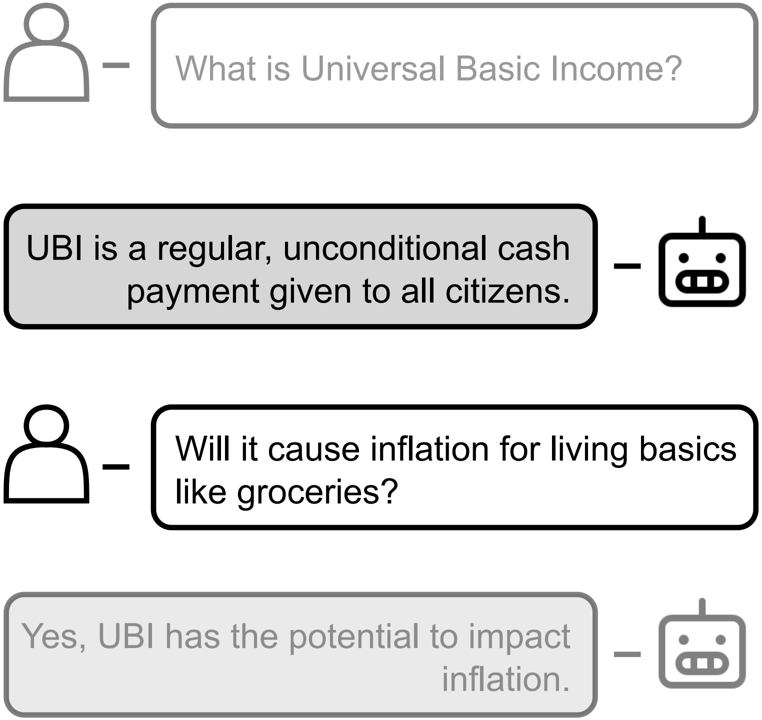
Follow-up Questions

Task: Given an informative textual response to a user's query, generate a question a user might ask based on the provided information.

RQ1. Are the generated questions similar to human questions?

RQ2. According to human judgments, are the generated questions appropriate follow-up questions?

RQ3. Can simple prompt modification result in simulation of different user profiles?



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Simulating Follow-up Questions in Conversational Search

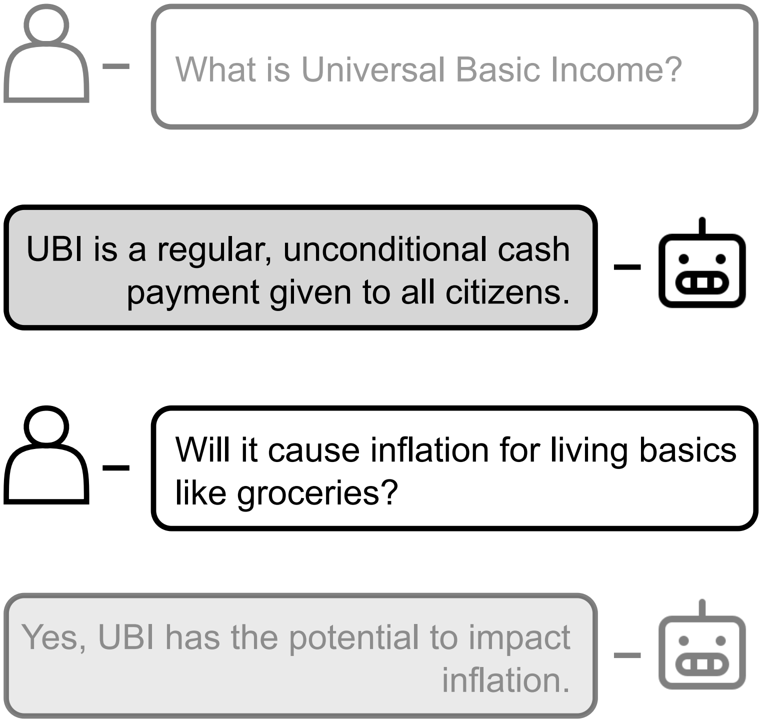
Follow-up Questions

Task: Given an informative textual response to a user's query, generate a question a user might ask based on the provided information.

RQ1. Are the generated questions similar to human questions?

RQ2. According to human judgments, are the generated questions appropriate follow-up questions?

RQ3. Can simple prompt modification result in simulation of different user profiles?



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Simulating Follow-up Questions in Conversational Search

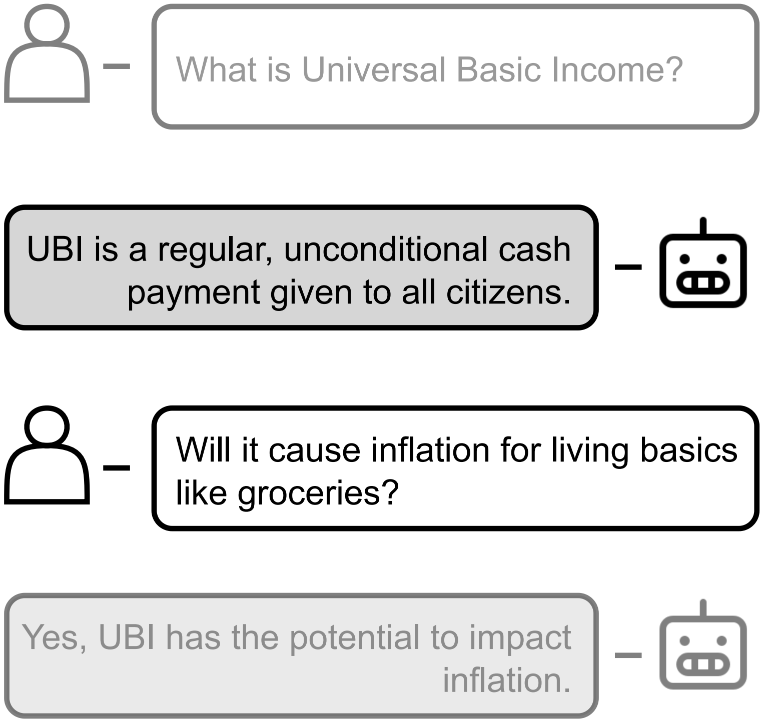
Follow-up Questions

Task: Given an informative textual response to a user's query, generate a question a user might ask based on the provided information.

RQ1. Are the generated questions similar to human questions?

RQ2. According to human judgments, are the generated questions appropriate follow-up questions?

RQ3. Can simple prompt modification result in simulation of different user profiles?



What is Universal Basic Income?

UBI is a regular, unconditional cash payment given to all citizens.

Will it cause inflation for living basics like groceries?

Yes, UBI has the potential to impact inflation.

Simulating Follow-up Questions in Conversational Search

Experimental Setup

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

Example of a prompt we employ to simulate a user asking questions.
Text with a red background is always adapted to the respective conversation

Datasets

1. TREC CAsT 2022
2. Webis-Nudged-Questions-23 (WNQ)

Models

1. Alpaca-7B
2. Llama-2-7B
3. Llama-2-13B
4. GPT-4

Approaches

1. Zero-shot prompting
2. Fine-tuning with LoRA
   except for GPT-4

Simulating Follow-up Questions in Conversational Search

Experimental Setup

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

Example of a prompt we employ to simulate a user asking questions.
Text with a red background is always adapted to the respective conversation

Datasets

1. TREC CAsT 2022
2. Webis-Nudged-Questions-23 (WNQ)

Models

1. Alpaca-7B
2. Llama-2-7B
3. Llama-2-13B
4. GPT-4

Approaches

1. Zero-shot prompting
2. Fine-tuning with LoRA
   except for GPT-4

Simulating Follow-up Questions in Conversational Search

Experimental Setup

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

Example of a prompt we employ to simulate a user asking questions.
Text with a red background is always adapted to the respective conversation

Datasets

1. TREC CAsT 2022
2. Webis-Nudged-Questions-23 (WNQ)

Models

1. Alpaca-7B
2. Llama-2-7B
3. Llama-2-13B
4. GPT-4

Approaches

1. Zero-shot prompting
2. Fine-tuning with LoRA
   except for GPT-4

Simulating Follow-up Questions in Conversational Search

Experimental Setup

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

Example of a prompt we employ to simulate a user asking questions.
Text with a red background is always adapted to the respective conversation

Datasets

1. TREC CAsT 2022
2. Webis-Nudged-Questions-23 (WNQ)

Models

1. Alpaca-7B
2. Llama-2-7B
3. Llama-2-13B
4. GPT-4

Approaches

1. Zero-shot prompting
2. Fine-tuning with LoRA
   except for GPT-4

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		BLEU		Sent.-BERT	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

- ❑ **BLEU** – lexical similarity
- ❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has ∼30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		BLEU		Sent.-BERT	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

❑ **BLEU** – lexical similarity

❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has ∼30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		BLEU		Sent.-BERT	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

- ❑ **BLEU** – lexical similarity
- ❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has ∼30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		BLEU		Sent.-BERT	
Base	Tuning	CAsT	WNQ	CAsT	WNQ
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

❑ **BLEU** – lexical similarity

❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has ~30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		**BLEU**		**Sent.-BERT**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

❑ **BLEU** – lexical similarity

❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has $\sim$30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		BLEU		Sent.-BERT	
Base	Tuning	CAsT	WNQ	CAsT	WNQ
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

- ❑ **BLEU** – lexical similarity
- ❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has ∼30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ1. Are the model generated questions similar to human generated questions?

Model		**BLEU**		**Sent.-BERT**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	0.02	0.11	0.22	0.68
Alpaca-7B	none	0.03	0.14	**0.23**	0.70
Alpaca-7B	CAsT	0.03	0.08	0.20	0.46
Alpaca-7B	WNQ	0.03	0.13	0.22	0.66
Llama2-7B	none	0.03	0.18	0.18	0.63
Llama2-7B	CAsT	**0.04**	0.09	0.19	0.45
Llama2-7B	WNQ	0.03	0.21	0.20	**0.71**
Llama2-13B	none	0.03	0.19	0.21	0.66
Llama2-13B	CAsT	**0.04**	0.07	0.20	0.41
Llama2-13B	WNQ	0.03	0.22	0.20	0.70

❑ **BLEU** – lexical similarity

❑ **Sentence-BERT** – semantic similarity

* All scores are higher for WNQ dataset, as it has ∼30 questions per response (vs mostly 1 in TREC CAsT).

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		**Valid**		**Related**		**Informative**		**Specific**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		**Valid**		**Related**		**Informative**		**Specific**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		**Valid**		**Related**		**Informative**		**Specific**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		**Valid**		**Related**		**Informative**		**Specific**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		Valid		Related		**Informative**		Specific	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		Valid		Related		**Informative**		Specific	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		**Valid**		**Related**		**Informative**		**Specific**	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		Valid		Related		Informative		Specific	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		Valid		Related		Informative		Specific	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ2. According to human judgments, are the generated questions appropriate?

Model		Valid		Related		Informative		Specific	
Base	**Tuning**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**	**CAsT**	**WNQ**
GPT-4	none	**0.98**	0.97	0.97	**0.97**	**0.84**	**0.87**	**0.84**	**0.87**
Alpaca-7B	none	0.93	0.97	0.93	0.93	0.73	0.63	0.72	0.63
Alpaca-7B	CAsT	0.92	0.87	0.85	0.80	0.80	0.80	0.72	0.70
Alpaca-7B	WNQ	0.96	0.77	0.94	0.77	0.75	0.67	0.75	0.67
Llama2-7B	none	0.92	0.80	0.84	0.77	0.60	0.50	0.57	0.47
Llama2-7B	CAsT	0.94	0.93	0.84	0.70	0.76	0.57	0.73	0.43
Llama2-7B	WNQ	0.96	**1.00**	0.94	0.93	0.65	0.63	0.65	0.63
Llama2-13B	none	0.90	0.93	0.88	0.90	0.57	0.50	0.52	0.43
Llama2-13B	CAsT	0.87	0.90	0.79	0.77	0.71	0.73	0.63	0.57
Llama2-13B	WNQ	0.94	0.97	0.89	0.93	0.58	0.60	0.58	0.57
Original questions	-	0.95	0.60	0.91	0.50	0.87	0.40	0.77	0.40

Ratio of simulated questions judged as valid, related, informative, and specific.
Each judgment implies the judgments to its left. Highest ratio in each column marked bold.

Simulating Follow-up Questions in Conversational Search

RQ3. Can simple prompt modification result in simulation of different user profiles?

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. You are a
[savvy/naive] user. You ask [elaborate/simple] questions about
the [implications/reasons] of what was being said. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

User is:

1. Savvy (asking elaborate questions)
2. Naive (asking simple questions)

Focusing on:

1. Implications
2. Reasons

Simulating Follow-up Questions in Conversational Search

RQ3. Can simple prompt modification result in simulation of different user profiles?

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. You are a
[savvy/naive] user. You ask [elaborate/simple] questions about
the [implications/reasons] of what was being said. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

User is:

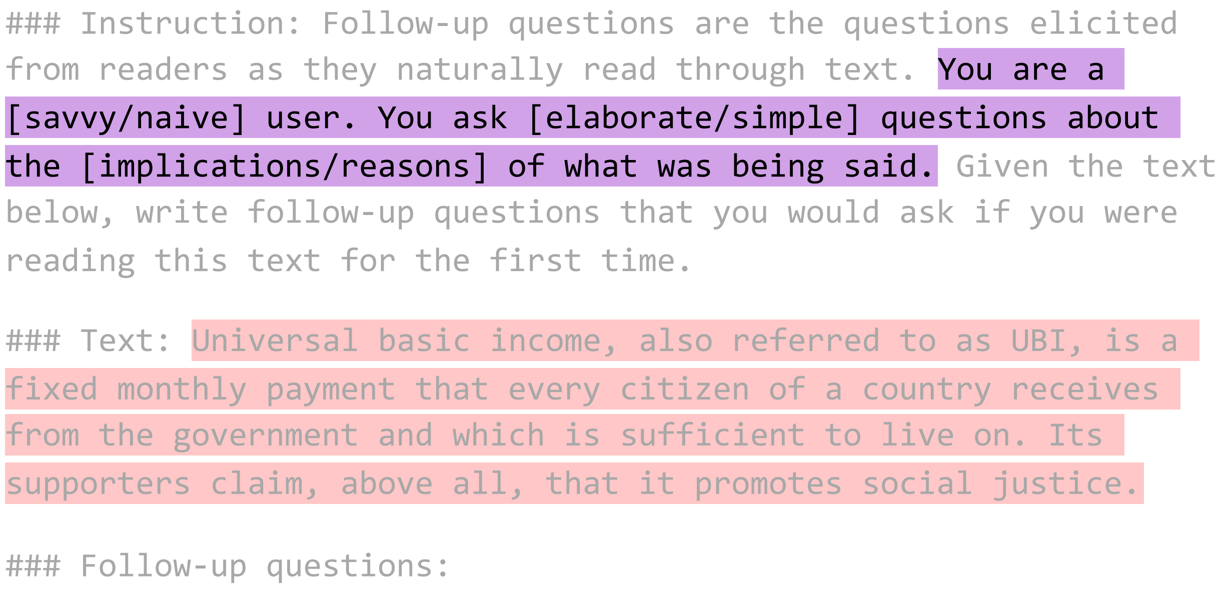
1. Savvy (asking elaborate questions)
2. Naive (asking simple questions)

Focusing on:

1. Implications
2. Reasons

Simulating Follow-up Questions in Conversational Search

RQ3. Can simple prompt modification result in simulation of different user profiles?



```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. You are a
[savvy/naive] user. You ask [elaborate/simple] questions about
the [implications/reasons] of what was being said. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

User is:

1. Savvy (asking elaborate questions)
2. Naive (asking simple questions)

Focusing on:

1. Implications
2. Reasons

Simulating Follow-up Questions in Conversational Search

RQ3. Can simple prompt modification result in simulation of different user profiles?

```
### Instruction: Follow-up questions are the questions elicited
from readers as they naturally read through text. You are a
[savvy/naive] user. You ask [elaborate/simple] questions about
the [implications/reasons] of what was being said. Given the text
below, write follow-up questions that you would ask if you were
reading this text for the first time.

### Text: Universal basic income, also referred to as UBI, is a
fixed monthly payment that every citizen of a country receives
from the government and which is sufficient to live on. Its
supporters claim, above all, that it promotes social justice.

### Follow-up questions:
```

GPT-4 examples:

1. How does the idea of UBI align with the core principles of capitalism and communism?
2. How does the government fund the Universal Basic Income?
3. Does the idea of universal basic income have any historical or cultural roots?
4. How is the amount of Universal Basic Income determined?

Simulating Follow-up Questions in Conversational Search

Conclusion

- ❑ Generated questions are semantically similar to human questions.
- ❑ Generated questions are valid, relevant, informative, and specific.
- ❑ Small prompt variations only minimally affect simulated user question traits.

Simulating Follow-up Questions in Conversational Search

Conclusion

- Generated questions are semantically similar to human questions.
- Generated questions are valid, relevant, informative, and specific.
- Small prompt variations only minimally affect simulated user question traits.

Code and Data

https://github.com/webis-de/ECIR-24

Simulating Follow-up Questions in Conversational Search

Conclusion

- Generated questions are semantically similar to human questions.
- Generated questions are valid, relevant, informative, and specific.
- Small prompt variations only minimally affect simulated user question traits.

Code and Data

https://github.com/webis-de/ECIR-24

Thank you!