Flipping The Dialogue:
Training and Evaluating User Language Models

Tarek Naous

Philippe Laban

Wei Xu

Jennifer Neville

🤗 https://huggingface.co/microsoft/UserLM-8b

ICLR International Conference On Learning Representations

Microsoft

Georgia Tech

You must act as a real user and make requests as a user would to an AI assistant. You should be brief and potentially make [...]

The task you are trying to accomplish as a user is: **getting information about strategies for losing weight** and the impact of certain drugs or medications on weight gain.

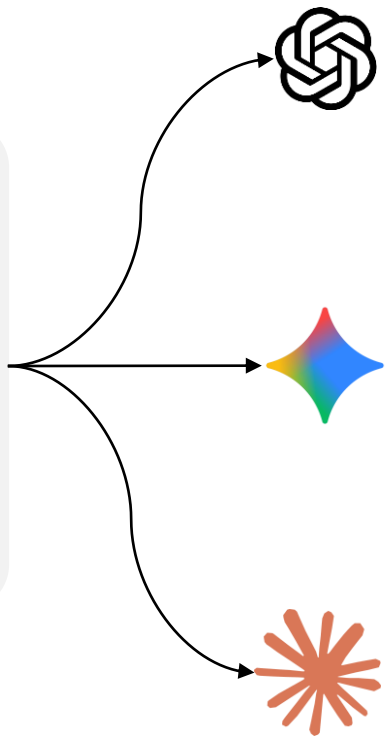
You must act as a real user and make requests as a user would to an AI assistant. You should be brief and potentially make [...]

The task you are trying to accomplish as a user is: **getting information about strategies for losing weight** and the impact of certain drugs or medications on weight gain.

hey, looking for tips on losing weight + how certain meds/drugs affect weight gain (like antidepressants, birth control, etc). any info?

You must act as a real user and make requests as a user would to an AI assistant. You should be brief and potentially make [...]

The task you are trying to accomplish as a user is: **getting information about strategies for losing weight** and the impact of certain drugs or medications on weight gain.



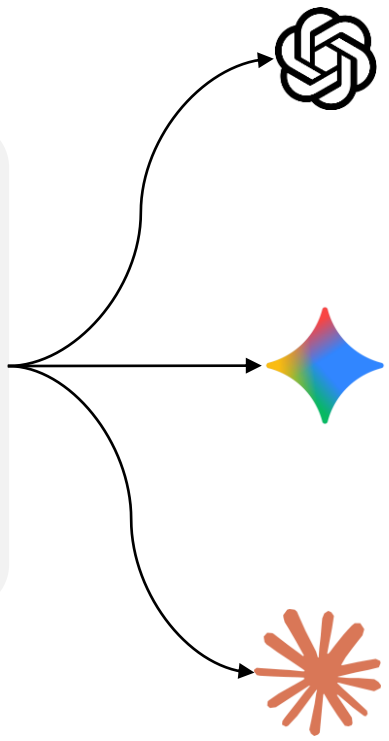
hey, looking for tips on losing weight + how certain meds/drugs affect weight gain (like antidepressants, birth control, etc). any info?

hey can you tell me about ozempic and other weight loss drugs? also what are some good strategies for losing weight in general

hey can u help me out? i've been trying to lose some weight but i feel like im stuck. also i started some new meds recently and i swear im gaining weight faster now.

You must act as a real user and make requests as a user would to an AI assistant. You should be brief and potentially make [...]

The task you are trying to accomplish as a user is: **getting information about strategies for losing weight** and the impact of certain drugs or medications on weight gain.



hey, looking for tips on losing weight + how certain meds/drugs affect weight gain (like antidepressants, birth control, etc). any info?

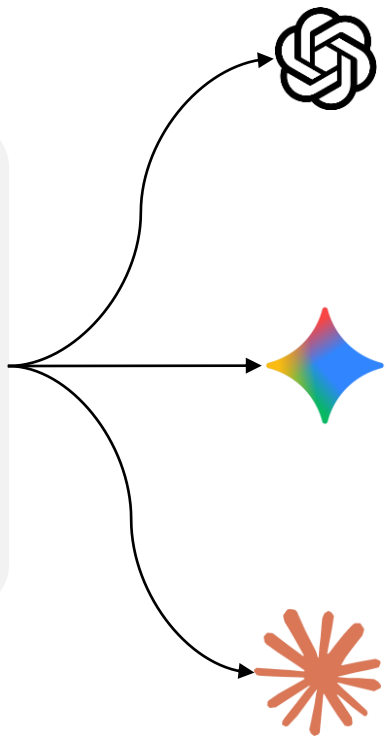
hey can you tell me about ozempic and other weight loss drugs? also what are some good strategies for losing weight in general

hey can u help me out? i've been trying to lose some weight but i feel like im stuck. also i started some new meds recently and i swear im gaining weight faster now.

Real User

You must act as a real user and make requests as a user would to an AI assistant. You should be brief and potentially make [...]

The task you are trying to accomplish as a user is: **getting information about strategies for losing weight** and the impact of certain drugs or medications on weight gain.



hey, looking for tips on losing weight + how certain meds/drugs affect weight gain (like antidepressants, birth control, etc). any info?

hey can you tell me about ozempic and other weight loss drugs? also what are some good strategies for losing weight in general

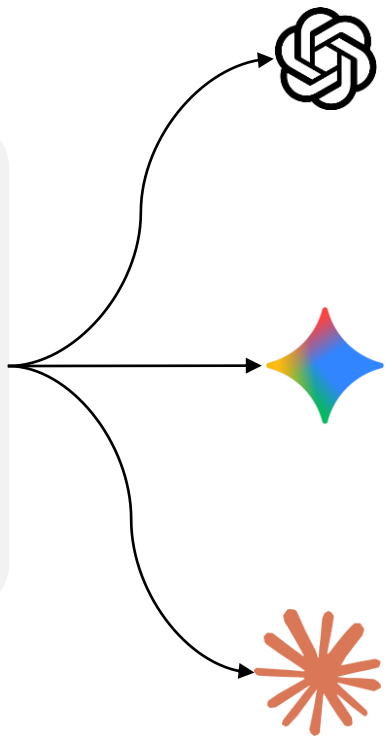
hey can u help me out? i've been trying to lose some weight but i feel like im stuck. also i started some new meds recently and i swear im gaining weight faster now.

Real User

does allertine cause weight gain

You must act as a real user and make requests as a user would to an AI assistant. You should be brief and potentially make [...]

The task you are trying to accomplish as a user is: **getting information about strategies for losing weight** and the impact of certain drugs or medications on weight gain.



hey, looking for tips on losing weight + how certain meds/drugs affect weight gain (like antidepressants, birth control, etc). any info?

hey can you tell me about ozempic and other weight loss drugs? also what are some good strategies for losing weight in general

hey can u help me out? i've been trying to lose some weight but i feel like im stuck. also i started some new meds recently and i swear im gaining weight faster now.

Real User does allertine cause weight gain

The inherent assistant role of language models makes them poor user simulators

SIMULATING USERS IN CONVERSATIONS …

User Intent: Write a Python function: given an array of integers, sort ones between 1 and 9 inclusive, reverse the array, and replace digits by their name from "One", "Two", "Three", etc.

SIMULATING USERS IN CONVERSATIONS …

… USING AN ASSISTANT LANGUAGE MODEL

User Intent: Write a Python function: given an array of integers, sort ones between 1 and 9 inclusive, reverse the array, and replace digits by their name from "One", "Two", "Three", etc.

GPT-4o

Turn digits into names in a list

...

I want to sort the numbers if they're between 1 and 9

...

I want to flip the list after that

...

I want to use names for numbers from One to Nine

Simple & Direct User Turns → ✔ Success

Benchmark Score: 74.6 ↑

SIMULATING USERS IN CONVERSATIONS …

… USING AN ASSISTANT LANGUAGE MODEL

User Intent: Write a Python function: given an array of integers, sort ones between 1 and 9 inclusive, reverse the array, and replace digits by their name from "One", "Two", "Three", etc.

GPT-4o

Turn digits into names in a list

...

I want to sort the numbers if they're between 1 and 9

...

I want to flip the list after that

...

I want to use names for numbers from One to Nine

Simple & Direct User Turns → ✔ Success

Benchmark Score: 74.6 ↑

… USING A USER LANGUAGE MODEL

UserLM-8b

ignore any number that is not in the 1-to-9 range

...

now add sorting into it

...

now reverse the sorted list

...

translate each remaining digit into its capitalized English name

Nuanced & Indirect User Turns → ✗ Failure

Benchmark Score: 57.4 ↓

SIMULATING USERS IN CONVERSATIONS …

… USING AN ASSISTANT LANGUAGE MODEL

… USING A USER LANGUAGE MODEL

User Intent: Write a Python function: given an array of integers, sort ones between 1 and 9 inclusive, reverse the array, and replace digits by their name from "One", "Two", "Three", etc.

GPT-4o

- Turn digits into names in a list
- …
- I want to sort the numbers if they're between 1 and 9
- …
- I want to flip the list after that
- …
- I want to use names for numbers from One to Nine

Simple & Direct User Turns → ✔ Success

Benchmark Score: 74.6 ↑

Simulator Capability

GPT-4o *(Prompted)*	Real Users	UserLM-8b *(Trained)*
Diversity (↑)		
74.4	94.01	94.5
Dialogue Termination (↑)		
1.3	100	63.5
Intent Decomposition (↓)		
7.7	1.68	2.7
Naturalness (↑)		
3.3	90.15	80.2
Robustness (↑)		
54.9	100	94.3

UserLM-8b

- ignore any number that is not in the 1-to-9 range
- …
- now add sorting into it
- …
- now reverse the sorted list
- …
- translate each remaining digit into its capitalized English name

Nuanced & Indirect User Turns → ✗ Failure

Benchmark Score: 57.4 ↓

Flipping The Dialogue

1 Training User LMs

How do we train a user language model?

User-Assistant Conversations

In-the-wild user-assistant conversations

CONVERSATION 1

does allertine cause weight gain

As an AI language model, I don't have personal experiences. But according to ..

do antihistamines cause weight gain

Most antihistamines do not cause weight gain as a side effect. However ..

how much weight can i lose by cutting sugar for 40 days

The amount of weight you can lose …

CONVERSATION N-1

…

CONVERSATION N

Is einstein determinism challenged by quantum mechanics?

Quantum mechanics challenges Einstein's determinism, which held that …

What did Hawkins say?

It seems you are referring to Stephen Hawking, the …

User-Assistant Conversations

In-the-wild user-assistant conversations

CONVERSATION 1

does allertine cause weight gain

As an AI language model, I don't have personal experiences. But according to ..

do antihistamines cause weight gain

Most antihistamines do not cause weight gain as a side effect. However ..

how much weight can i lose by cutting sugar for 40 days

The amount of weight you can lose …

CONVERSATION N-1

...

CONVERSATION N

Is einstein determinism challenged by quantum mechanics?

Quantum mechanics challenges Einstein's determinism, which held that …

What did Hawkins say?

It seems you are referring to Stephen Hawking, the …

User Intent Generation

Generate generic user intents for conversations

INTENT 1

You are a user talking with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.

INTENT N-1

You are a user talking with an assistant language model to …

INTENT N

You are a user talking with an assistant language model to understand how quantum Mechanics challenges Einstein's determinism and get the perspective of other scientists.

User-Assistant Conversations

In-the-wild user-assistant conversations

CONVERSATION 1

does allertine cause weight gain

As an AI language model, I don't have personal experiences. But according to ..

do antihistamines cause weight gain

Most antihistamines do not cause weight gain as a side effect. However ..

how much weight can i lose by cutting sugar for 40 days

The amount of weight you can lose …

CONVERSATION N-1

...

CONVERSATION N

Is einstein determinism challenged by quantum mechanics?

Quantum mechanics challenges Einstein's determinism, which held that …

What did Hawkins say?

It seems you are referring to Stephen Hawking, the …

User Intent Generation

Generate generic user intents for conversations

1 conversation (**3 turns**) + intent yields **4 samples**

INTENT 1

You are a user talking with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.

INTENT N-1

You are a user talking with an assistant language model to …

INTENT N

You are a user talking with an assistant language model to understand how quantum Mechanics challenges Einstein's determinism and get the perspective of other scientists.

User Language Modeling

Flip the dialogue turns to create training samples for a User LM

User-Assistant Conversations

In-the-wild user-assistant conversations

CONVERSATION 1

does allertine cause weight gain

As an AI language model, I don't have personal experiences. But according to ..

do antihistamines cause weight gain

Most antihistamines do not cause weight gain as a side effect. However ..

how much weight can i lose by cutting sugar for 40 days

The amount of weight you can lose …

CONVERSATION N-1

...

CONVERSATION N

Is einstein determinism challenged by quantum mechanics?

Quantum mechanics challenges Einstein's determinism, which held that …

What did Hawkins say?

It seems you are referring to Stephen Hawking, the …

User Intent Generation

Generate generic user intents for conversations

1 conversation (**3 turns**) + intent yields **4 samples**

INTENT 1

You are a user talking with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.

INTENT N-1

You are a user talking with an assistant language model to …

INTENT N

You are a user talking with an assistant language model to understand how quantum Mechanics challenges Einstein's determinism and get the perspective of other scientists.

User Language Modeling

Flip the dialogue turns to create training samples for a User LM

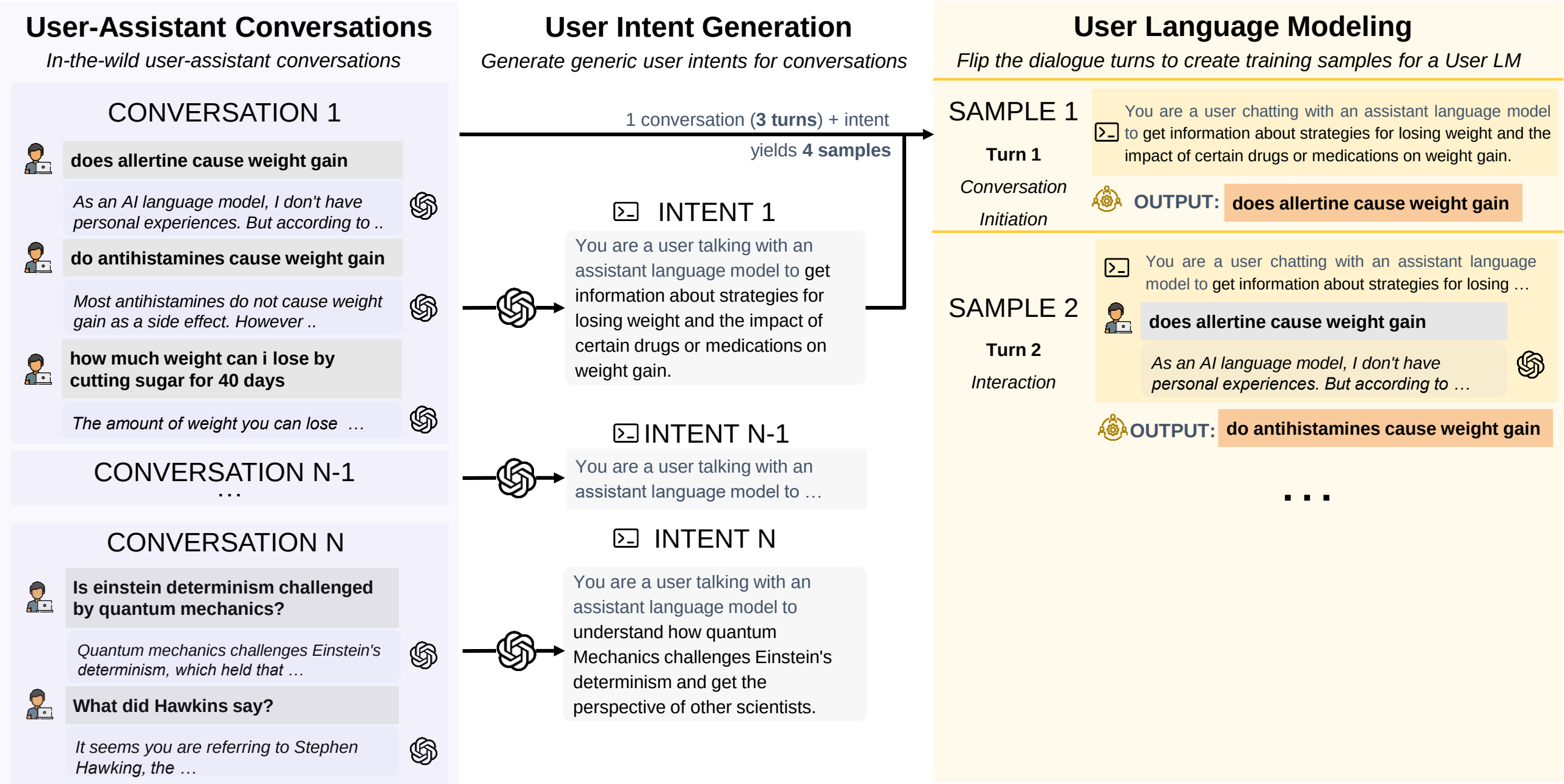
SAMPLE 1

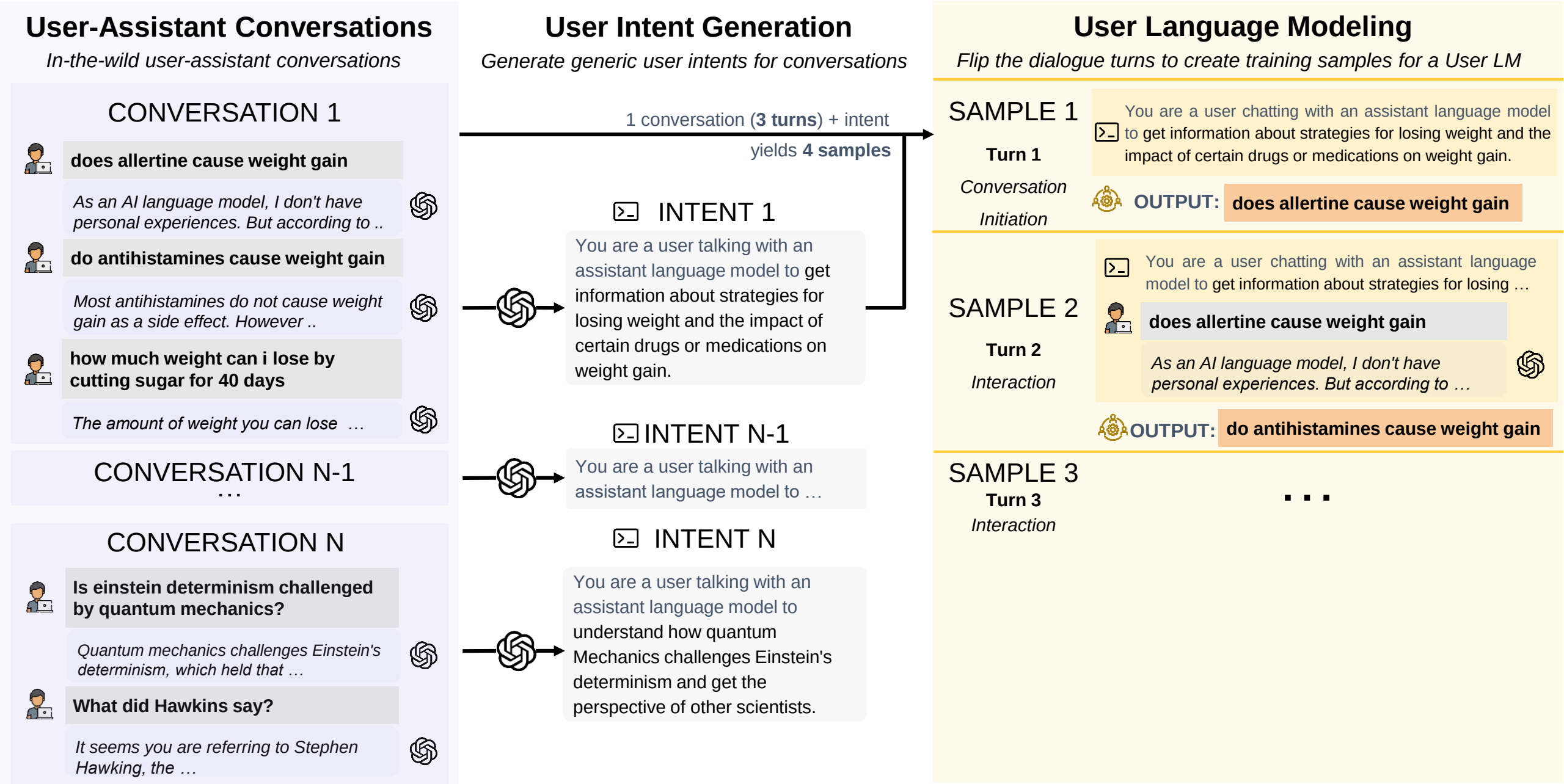
Turn 1

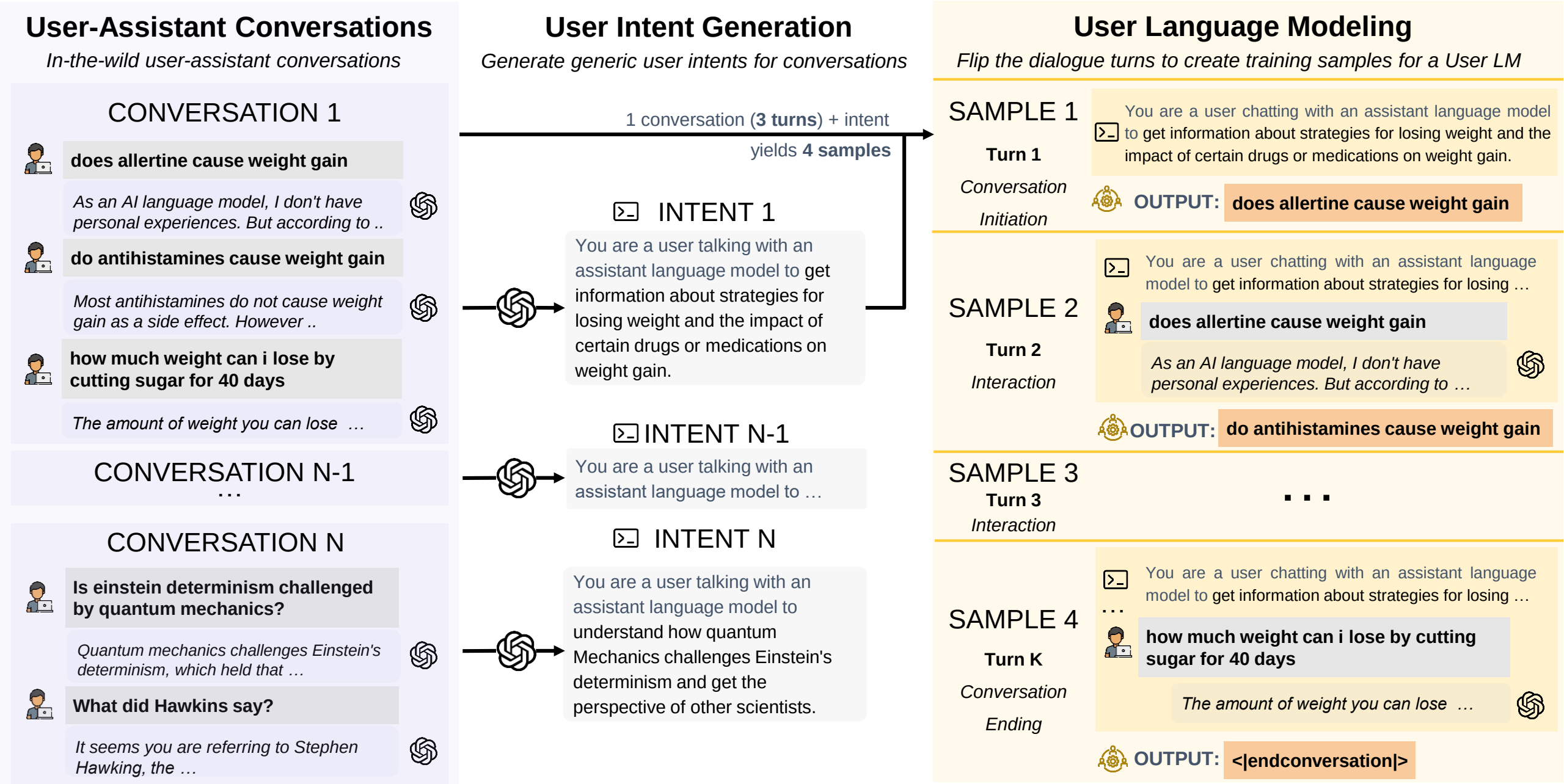
Conversation Initiation

You are a user chatting with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.

OUTPUT: **does allertine cause weight gain**


User-Assistant Conversations
In-the-wild user-assistant conversations
CONVERSATION 1
does allertine cause weight gain
As an AI language model, I don't have personal experiences. But according to ..
do antihistamines cause weight gain
Most antihistamines do not cause weight gain as a side effect. However ..
how much weight can i lose by cutting sugar for 40 days
The amount of weight you can lose …
CONVERSATION N-1
…
CONVERSATION N
Is einstein determinism challenged by quantum mechanics?
Quantum mechanics challenges Einstein's determinism, which held that …
What did Hawkins say?
It seems you are referring to Stephen Hawking, the …
User Intent Generation
Generate generic user intents for conversations
1 conversation (3 turns) + intent yields 4 samples
INTENT 1
You are a user talking with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.
INTENT N-1
You are a user talking with an assistant language model to …
INTENT N
You are a user talking with an assistant language model to understand how quantum Mechanics challenges Einstein's determinism and get the perspective of other scientists.
User Language Modeling
Flip the dialogue turns to create training samples for a User LM
SAMPLE 1
Turn 1
Conversation Initiation
You are a user chatting with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.
OUTPUT: does allertine cause weight gain
SAMPLE 2
Turn 2
Interaction
You are a user chatting with an assistant language model to get information about strategies for losing …
does allertine cause weight gain
As an AI language model, I don't have personal experiences. But according to …
OUTPUT: do antihistamines cause weight gain
. . .


User-Assistant Conversations
In-the-wild user-assistant conversations
CONVERSATION 1
does allertine cause weight gain
As an AI language model, I don't have personal experiences. But according to ..
do antihistamines cause weight gain
Most antihistamines do not cause weight gain as a side effect. However ..
how much weight can i lose by cutting sugar for 40 days
The amount of weight you can lose …
CONVERSATION N-1
…
CONVERSATION N
Is einstein determinism challenged by quantum mechanics?
Quantum mechanics challenges Einstein's determinism, which held that …
What did Hawkins say?
It seems you are referring to Stephen Hawking, the …
User Intent Generation
Generate generic user intents for conversations
1 conversation (3 turns) + intent yields 4 samples
INTENT 1
You are a user talking with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.
INTENT N-1
You are a user talking with an assistant language model to …
INTENT N
You are a user talking with an assistant language model to understand how quantum Mechanics challenges Einstein's determinism and get the perspective of other scientists.
User Language Modeling
Flip the dialogue turns to create training samples for a User LM
SAMPLE 1
Turn 1
Conversation Initiation
You are a user chatting with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.
OUTPUT: does allertine cause weight gain
SAMPLE 2
Turn 2
Interaction
You are a user chatting with an assistant language model to get information about strategies for losing …
does allertine cause weight gain
As an AI language model, I don't have personal experiences. But according to …
OUTPUT: do antihistamines cause weight gain
SAMPLE 3
Turn 3
Interaction
▪ ▪ ▪


User-Assistant Conversations
In-the-wild user-assistant conversations
CONVERSATION 1
does allertine cause weight gain
As an AI language model, I don't have personal experiences. But according to ..
do antihistamines cause weight gain
Most antihistamines do not cause weight gain as a side effect. However ..
how much weight can i lose by cutting sugar for 40 days
The amount of weight you can lose …
CONVERSATION N-1
…
CONVERSATION N
Is einstein determinism challenged by quantum mechanics?
Quantum mechanics challenges Einstein's determinism, which held that …
What did Hawkins say?
It seems you are referring to Stephen Hawking, the …
User Intent Generation
Generate generic user intents for conversations
1 conversation (3 turns) + intent
yields 4 samples
INTENT 1
You are a user talking with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.
INTENT N-1
You are a user talking with an assistant language model to …
INTENT N
You are a user talking with an assistant language model to understand how quantum Mechanics challenges Einstein's determinism and get the perspective of other scientists.
User Language Modeling
Flip the dialogue turns to create training samples for a User LM
SAMPLE 1
Turn 1
Conversation Initiation
You are a user chatting with an assistant language model to get information about strategies for losing weight and the impact of certain drugs or medications on weight gain.
OUTPUT: does allertine cause weight gain
SAMPLE 2
Turn 2
Interaction
You are a user chatting with an assistant language model to get information about strategies for losing …
does allertine cause weight gain
As an AI language model, I don't have personal experiences. But according to …
OUTPUT: do antihistamines cause weight gain
SAMPLE 3
Turn 3
Interaction
. . .
SAMPLE 4
Turn K
Conversation Ending
You are a user chatting with an assistant language model to get information about strategies for losing …
how much weight can i lose by cutting sugar for 40 days
The amount of weight you can lose …
OUTPUT: <|endconversation|>

Training Setup and Details

Data: 384,336 conversations from WildChat after filtering/deduplication (1,047,930 training samples)

Intent Generation: generated for each conversation using few-shot prompting with GPT-4o

Models: We train UserLM-1b and UserLM-8b starting from Llama3-8b-Base and Llama3.2-1b-Base

Training Details: max seq length of 2048 tokens, batch size of 1024 samples, learning rate of 2e-5

Distributional Alignment with Human Utterances

Perplexity on user utterances in two setups:

	WildChat
User Simulator	
Llama3.2-1b-Instruct	
Llama3-8b-Instruct	
GPT-4o-mini	
GPT-4o	
USP-8b	
UserLM-1b	
UserLM-8b	

Distributional Alignment with Human Utterances

Perplexity on user utterances in two setups:

Without intent conditioning

	WildChat
User Simulator	(↓)
Llama3.2-1b-Instruct	37.68
Llama3-8b-Instruct	98.29
GPT-4o-mini	26.08
GPT-4o	26.19
USP-8b	32.08
UserLM-1b	8.30
UserLM-8b	**5.60**

Distributional Alignment with Human Utterances

Perplexity on user utterances in two setups:

Without intent conditioning

With intent conditioning

	WildChat	
User Simulator	(↓)	(↓)
Llama3.2-1b-Instruct	37.68	29.09
Llama3-8b-Instruct	98.29	48.13
GPT-4o-mini	26.08	16.08
GPT-4o	26.19	21.40
USP-8b	32.08	21.78
UserLM-1b	8.30	7.78
UserLM-8b	**5.60**	**4.33**

Distributional Alignment with Human Utterances

Perplexity on user utterances in two setups:

Without intent conditioning

With intent conditioning

	in-domain		out-of-domain	
	WildChat		**PRISM**	
User Simulator	(↓)	(↓)	(↓)	(↓)
Llama3.2-1b-Instruct	37.68	29.09	84.00	53.54
Llama3-8b-Instruct	98.29	48.13	89.98	40.86
GPT-4o-mini	26.08	16.08	35.02	20.80
GPT-4o	26.19	21.40	40.25	36.29
USP-8b	32.08	21.78	50.91	30.16
UserLM-1b	8.30	7.78	18.58	10.33
UserLM-8b	**5.60**	**4.33**	**14.92**	**7.42**

Distributional Alignment with Human Utterances

Perplexity on user utterances in two setups:

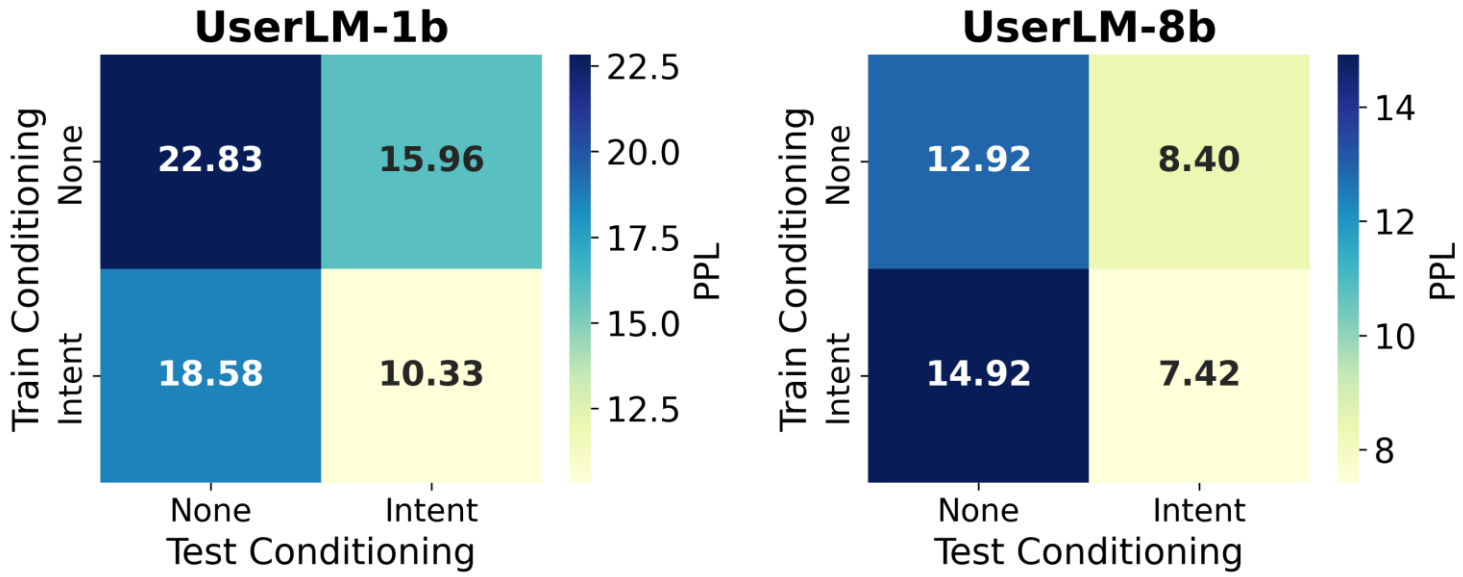
With intent conditioning

User LMs show better distributional alignment with user text
(*better with intent conditioning, generalize to out-of-domain data*)

	in-domain		out-of-domain	
	WildChat		PRISM	
User Simulator	(↓)	(↓)	(↓)	(↓)
GPT-4o	26.19	21.40	40.25	36.29
USP-8b	32.08	21.78	50.91	30.16
UserLM-1b	8.30	7.78	18.58	10.33
UserLM-8b	**5.60**	**4.33**	**14.92**	**7.42**

How important is conditioning on generic intent?

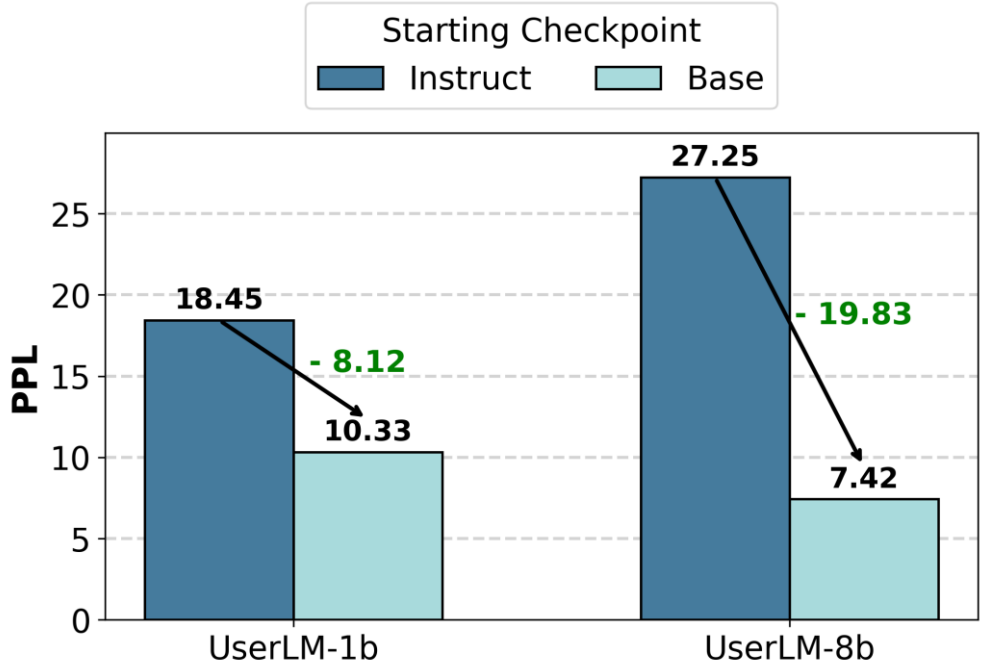
We train User LMs without any first-turn intent conditioning and compare their test-time performance



The biggest PPL gains are observed when models are trained with intent conditioning

Is starting from a base or instruction-tuned checkpoint better?

We compare our User LMs to ones we train starting from the instruct Llama models



Starting from a base checkpoint results in better PPL than ones tuned for the assistant role

Flipping The Dialogue

1 **Training User LMs** *How do we train a user language model?*

2 **Intrinsic Evaluation** *How do we evaluate a user language model?*

Evaluating Multi-turn Interaction Properties

User Simulator
Llama3.2-1b-Instruct
Llama3-8b-Instruct
GPT-4o-mini
GPT-4o
USP-8b
UserLM-1b
UserLM-8b
Human *(estimate)*

Evaluating Multi-turn Interaction Properties

First-turn Diversity
Are first turn requests generated diverse?

User Simulator	(↑)
Llama3.2-1b-Instruct	81.36
Llama3-8b-Instruct	81.31
GPT-4o-mini	66.10
GPT-4o	74.42
USP-8b	94.37
UserLM-1b	90.90
UserLM-8b	**94.55**
Human *(estimate)*	94.01

Evaluating Multi-turn Interaction Properties

First-turn Diversity
Are first turn requests generated diverse?

Intent Decomposition
Does the simulator split its intent across turns?

User Simulator	(↑)	(↓)
Llama3.2-1b-Instruct	81.36	15.72
Llama3-8b-Instruct	81.31	23.95
GPT-4o-mini	66.10	9.66
GPT-4o	74.42	7.68
USP-8b	94.37	6.33
UserLM-1b	90.90	3.07
UserLM-8b	**94.55**	**2.69**
Human *(estimate)*	94.01	1.68

Evaluating Multi-turn Interaction Properties

First-turn Diversity
Are first turn requests generated diverse?

Intent Decomposition
Does the simulator split its intent across turns?

Dialogue Termination
Does the simulator end the conversation ?

User Simulator	(↑)	(↓)	(↑)
Llama3.2-1b-Instruct	81.36	15.72	3.47
Llama3-8b-Instruct	81.31	23.95	3.51
GPT-4o-mini	66.10	9.66	15.31
GPT-4o	74.42	7.68	1.38
USP-8b	94.37	6.33	21.31
UserLM-1b	90.90	3.07	56.83
UserLM-8b	**94.55**	**2.69**	**63.54**
Human *(estimate)*	94.01	1.68	—

Evaluating Multi-turn Interaction Properties

First-turn Diversity
Are first turn requests generated diverse?

Dialogue Termination
Does the simulator end the conversation ?

User LMs align better with real users properties
(*requests are diverse, intents are split across turns, dialogues end*)

User Simulator	(↑)	(↓)	(↑)
Llama3.2-1b-Instruct	81.36	15.72	3.47
Llama3-8b-Instruct	81.31	23.95	3.51
UserLM-1b	90.90	3.07	56.83
UserLM-8b	**94.55**	**2.69**	**63.54**
Human *(estimate)*	94.01	1.68	—

Evaluating Simulator Robustness

User Simulator
Llama3.2-1b-Instruct
Llama3-8b-Instruct
GPT-4o-mini
GPT-4o
USP-8b
UserLM-1b
UserLM-8b
Human *(estimate)*

Evaluating Simulator Robustness

Naturalness
Do simulator utterances resemble natural user text?

User Simulator	(↑)
Llama3.2-1b-Instruct	0.14
Llama3-8b-Instruct	0.20
GPT-4o-mini	0.04
GPT-4o	3.31
USP-8b	77.73
UserLM-1b	78.96
UserLM-8b	**80.21**
Human *(estimate)*	90.15

Evaluating Simulator Robustness

Naturalness
Do simulator utterances resemble natural user text?

User Role Adherence
Does the simulator stick to its role as the user?

User Simulator	(↑)	(↑)
Llama3.2-1b-Instruct	0.14	77.55
Llama3-8b-Instruct	0.20	63.25
GPT-4o-mini	0.04	80.20
GPT-4o	3.31	38.85
USP-8b	77.73	**98.05**
UserLM-1b	78.96	91.30
UserLM-8b	**80.21**	93.95
Human *(estimate)*	90.15	—

Evaluating Simulator Robustness

Naturalness
Do simulator utterances resemble natural user text?

User Role Adherence
Does the simulator stick to its role as the user?

User Intent Adherence
Does the simulator stick to its user intent?

User Simulator	(↑)	(↑)	(↑)
Llama3.2-1b-Instruct	0.14	77.55	54.95
Llama3-8b-Instruct	0.20	63.25	78.05
GPT-4o-mini	0.04	80.20	88.70
GPT-4o	3.31	38.85	70.95
USP-8b	77.73	**98.05**	**97.55**
UserLM-1b	78.96	91.30	93.55
UserLM-8b	**80.21**	93.95	94.65
Human *(estimate)*	90.15	—	—

Evaluating Simulator Robustness

Naturalness
Do simulator utterances resemble natural user text?

User Intent Adherence
Does the simulator stick to its user intent?

User LMs are more robust simulators than assistant LMs
(*stick to their user role, do not change their intent easily*)

User Simulator	(↑)	(↑)	(↑)
Llama3.2-1b-Instruct	0.14	77.55	54.95
USP-8b	77.73	**98.05**	**97.55**
UserLM-1b	78.96	91.30	93.55
UserLM-8b	**80.21**	93.95	94.65
Human *(estimate)*	90.15	—	—

Flipping The Dialogue

1 **Training User LMs** — *How do we train a user language model?*

2 **Intrinsic Evaluation** — *How do we evaluate a user language model?*

3 Extrinsic Evaluation — *How do user language models help in practice?*

Extrinsic Eval Setup: Multi-Turn Evaluation of Assistants

We simulate multi-turn conversations between a GPT-4o assistant and UserLM-8b

Tasks: 65 task intents that involve users completing:

- Math word problems (*based on GSM8k* (Cobbe et al., 2021))
- Writing Python programs (*based on HumanEval* (Chen et al., 2021))

Baselines: We compare to simulators based on prompted GPT-4o and 4o-mini

Evaluation: We analyze various properties of the simulated conversations for each simulator and compare to how the performance of the assistant changes

	User Simulator		
	4o-mini	GPT-4o	UserLM-8b
	Intent Coverage		
Intent Coverage (%)	86.6	84.7	76.7

	User Simulator		
	4o-mini	GPT-4o	UserLM-8b
	Intent Coverage		
Intent Coverage (%)	86.6	84.7	76.7
	Information Diversity		
%Repeat Required	31.8	9.4	54.3
%Skip Non-Required	10.9	14.6	37.7
%Add Demands	9.5	1.1	43.8

UserLM-8b introduces more nuances to the conversation

	User Simulator		
	4o-mini	GPT-4o	UserLM-8b
	Intent Coverage		
Intent Coverage (%)	86.6	84.7	76.7
	Information Diversity		
%Repeat Required	31.8	9.4	54.3
%Skip Non-Required	10.9	14.6	37.7
%Add Demands	9.5	1.1	43.8
	Pace Diversity		
Turn Variance	0.9	0.6	2.8
Turn Range	3.7-5.7	4.0-5.4	2.1-6.7

UserLM-8b introduces more nuances to the conversation

UserLM-8b simulates more varied conversational paces

	User Simulator		
	4o-mini	GPT-4o	UserLM-8b
	Intent Coverage		
Intent Coverage (%)	86.6	84.7	76.7
	Information Diversity		
%Repeat Required	31.8	9.4	54.3
%Skip Non-Required	10.9	14.6	37.7
%Add Demands	9.5	1.1	43.8
	Pace Diversity		
Turn Variance	0.9	0.6	2.8
Turn Range	3.7-5.7	4.0-5.4	2.1-6.7
	Lexical Diversity		
Unigram Difference	0.43	0.40	0.71

UserLM-8b introduces more nuances to the conversation

UserLM-8b simulates more varied conversational paces

UserLM-8b uses more varied phrasing of requests

	User Simulator		
	4o-mini	GPT-4o	UserLM-8b
	Intent Coverage		
Intent Coverage (%)	86.6	84.7	76.7
	Information Diversity		
%Repeat Required	31.8	9.4	54.3
%Skip Non-Required	10.9	14.6	37.7
%Add Demands	9.5	1.1	43.8
	Pace Diversity		
Turn Variance	0.9	0.6	2.8
Turn Range	3.7-5.7	4.0-5.4	2.1-6.7
	Lexical Diversity		
Unigram Difference	0.43	0.40	0.71
	Assistant Diversity		
Assistant Score	73.2	74.6	**57.4**

UserLM-8b introduces more nuances to the conversation

UserLM-8b simulates more varied conversational paces

UserLM-8b uses more varied phrasing of requests

Better estimate of assistant performance

	User Simulator		
	4o-mini	GPT-4o	UserLM-8b
Intent Coverage			
Intent Coverage (%)	86.6	84.7	76.7
Information Diversity			
Turn Range	3.7-5.7	4.0-5.4	2.1-6.7
Lexical Diversity			
Unigram Difference	0.43	0.40	0.71
Assistant Diversity			
Assistant Score	73.2	74.6	**57.4**

Assistants struggle to handle user nuances in multi-turn settings
(UserLM-8b covers the necessary information to solve the tasks, but the nuances it introduces causes assistant failure)

UserLM-8b simulates more varied conversational paces

UserLM-8b uses more varied phrasing of requests

Better estimate of assistant performance

Where do we go from here?

Takeaways:

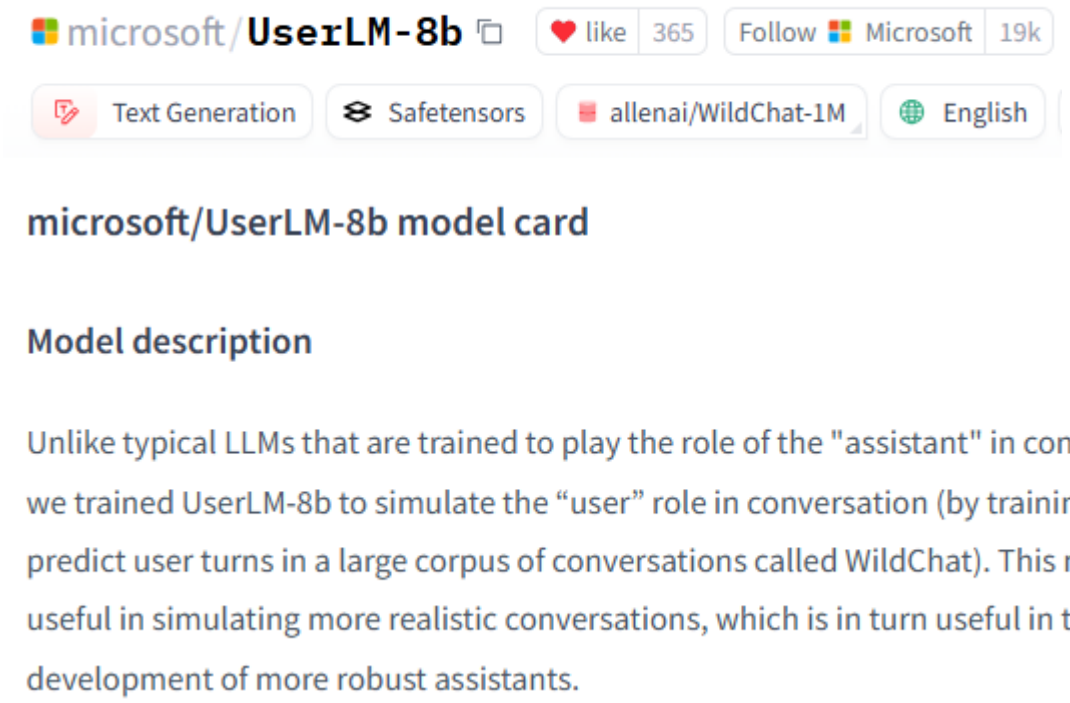
- The assistant role models are post-trained to play limits their capabilities for user simulation
- Language models should be trained to play distinct and opposing roles: ***user*** and ***assistant***
- User LMs will enable the development of robust assistant that are ready for real-world nuances

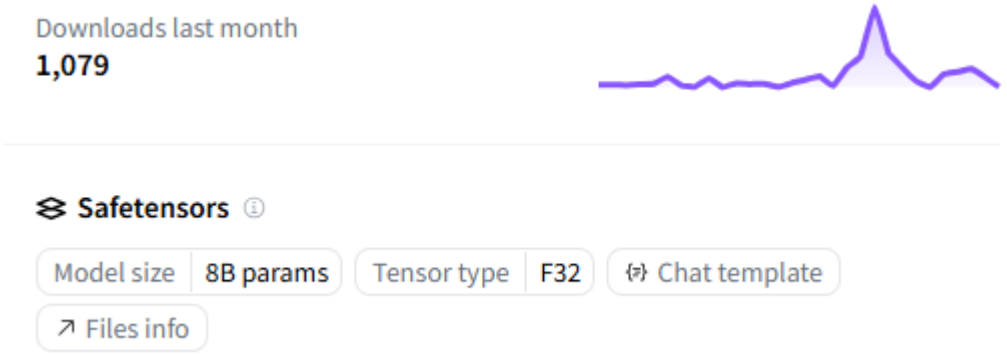
Notes for the future:

- More releases of better and bigger base models will be needed
- User LMs can serve purposes beyond multi-turn eval (synthetic data, user-centered judge, etc.)
- More personalized user LMs will be needed to simulate specific sub-populations

ขอบคุณ شكرا Merci 谢谢 धन्यवाद Asante Teşekkürler

ありがとう Gracias متشكرم நன்றி Obrigado Thank You





UserLM-8b is available at: 🤗 https://huggingface.co/microsoft/UserLM-8b

Feel free to follow up with me on 𝕏 @tareknaous