SocialLLM

Bae Sun Woo

April 9, 2026

Table of Contents

Position: LLM Social Simulations Are a Promising Research Method - Five key challenges.

Table 1: LLM social simulations must address five key challenges.

Challenge	Description	Promising Directions
Diversity	Generic and stereotypical outputs that lack human diversity	Inject humanlike variation in training, tuning, or inference (e.g., interview-based prompting, steering vectors)
Bias	Systematic inaccuracies when simulating particular human groups	Prompt with implicit demographic information; minimize accuracy-decreasing biases rather than all social biases
Sycophancy	Inaccuracies due to excessively user-pleasing outputs	Reduce the influence of instruction-tuning; instruct LLM to predict as an expert rather than roleplay a persona
Alienness	Superficially accurate results generated by non-humanlike mechanisms	Simulate latent features; iteratively conceptualize and evaluate; reassess as mechanistic interpretability advances
Generalization	Inaccuracies in out-of-distribution contexts, limiting scientific discovery	Simulate latent features; iteratively conceptualize and evaluate; reassess as generalization capabilities advance

Promising Direction 1: Prompting

	Explicit Demographics	Implicit Demographics	Distribution Elicitation (LLM-as-an-expert)
Example	"You are a 40-year-old Hispanic woman, conservative, income $50K"	"You are Maria Gonzalez from San Antonio, Texas"	"Predict what % of Americans would choose each option"/ ("You will be asked to predict how people respond to various messages")
Effect	Increase Diversity (Across-group diversity ↑)	Increase Diversity	Reduce Sycophancy
Problem	Increase Sterotype (Within-group diversity ↓)	Still has Sterotype	Still has sycophacy (If LLMs can understand what simulation researchers want)

Promising Direction 2: Steering Vectors

"The injection of variation directly into the embedding space via steering vectors"
(race, gender or political conservativeness)

Impacts on embedding space => reduced sycophancy

"Exciting potential for this approach but recommend caution in applied work pending further validation"

Promising Direction 3: Token Sampling

High Temperature => Flatten Distrubtion => High Diversity

However, low coherence

"Varying parameters for different LLM generations"

General Features : High Diversity(High T)
Specific Features : Lower Diversity(Low T)
(For self coherence)

Promising Direction 4: Training and Tuning

"Most contemporary models are optimized to be general-purpose assistants"

Training Entirely New Model ?  Difficult. Extremely large scales

Using Base model directly?  Without instruction tuned, predicting next sequence
+ Most SOTA models,
no public access to base model

"Fine-Tuning for instruction-tuned model"
=> reduce the influence of the instruction-tuning

Details in Next Paper

Finetuning LLMs for Human Behavior Prediction in Social Science

LLMs can simulate human behavior — but prompting alone has critical failures:

2–10× **Effect size overestimation**

LLMs exaggerate the magnitude of experimental manipulations

10–32% **Effect direction errors**

LLMs incorrectly predict the direction of significant effects

Flattened **Demographic variance loss**

LLMs compress variation across demographic subgroups

=> finetune LLMs directly on real human responses

Dataset: SOCSCI210 & Task Formulation

Dataset	Size			Feats.		Domain
	Source	Individuals	Total Data Points	D	I	
Psych-101 (Binz et al., 2024)	160 experiments	60,000	10,000,000	✗	✓	Psychology
SubPOP (Orlikowski et al., 2025)	3,362 questions	—	70,000	✓	✗	Public Opinion
E-commerce (Lu et al., 2025)	31,865 sessions	3,526	230,965	✗	✓	E-commerce
OpinionQA (Santurkar et al., 2023)	∼1,500 questions	—	90,000	✓	✗	Public Opinion
	50 questions	17,667	883,350	✓	✓	Big-5 Personality
Be.FM (Xie et al., 2025)	6 games	68,790	82,057	✓	✓	Behavioral Econ.
	2,703 abstracts	—	2,703	✗	✗	Behavioral Science
SocSci210 (Ours)	210 experiments	**400,491**	2,900,000	✓	✓	Social Sciences

Task Formulation

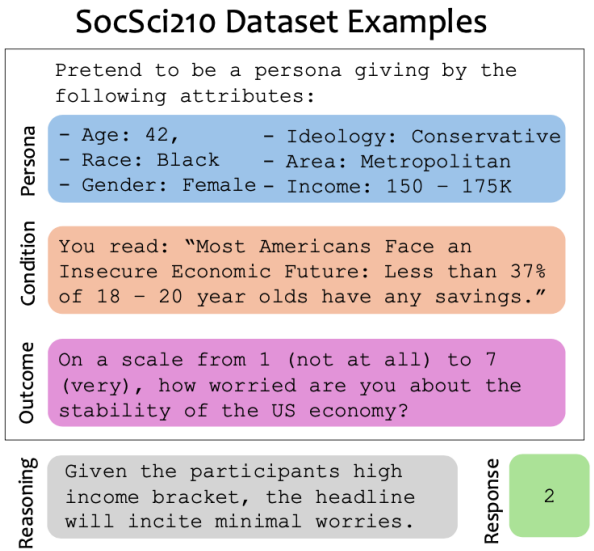
F (P, c, o) => r

P (Persona) Demographic attributes: age, race, ideology, income, ...

c (Condition) Experimental stimulus — the text the participant is exposed to

o (Outcome) Survey question — ordinal (1–7 scale) or binary (yes/no)

r (Response) Actual human answer recorded in the study — ground truth label



Finetuning Methods

SFT

F(P,c,o) => r

- Directly predicts actual response value

- **Best for distribution alignment**

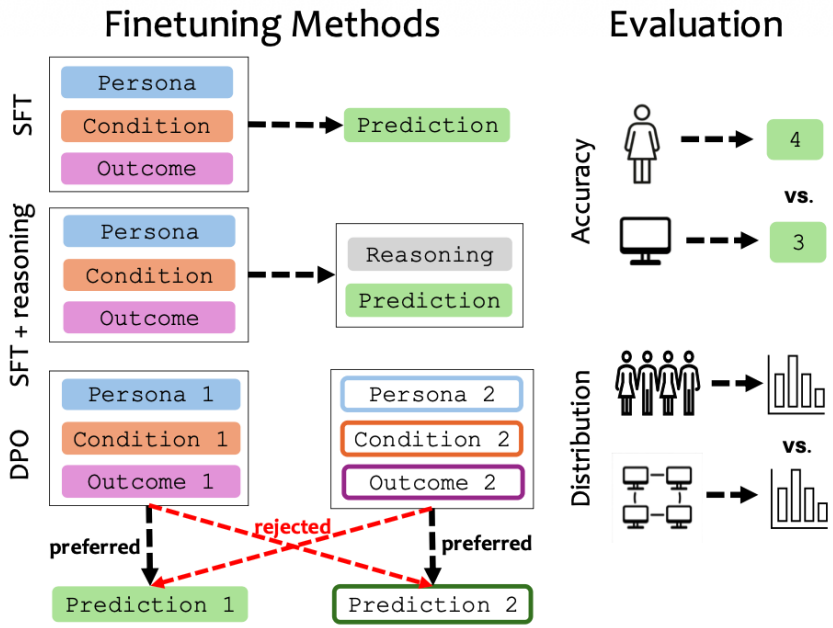
  ∵ Train all pairs (P,c,o)

=> Capture the shape of Distribution

SFT + Reasoning

F(P,c,o) => trace + r

- GPT-4o-mini generates oracle reasoning traces



DPO

$(P_{pos}$,c,o,$r_{pos})$ > $(P_{pos}$,c,o, $r_{neg})$

- To enhance the model's ability to differentiate responses based on variations in P, c and o

1. $(P_{pos}$,c,o,$r_{pos})$
2. Same (c,o) but different P, $(P_{neg}$,c,o,$r_{neg})$
3. Make a rejected pair, $(P_{pos}$,c,o,$r_{neg})$
4. DPO **$(P_{pos}$,c,o,$r_{pos})$ > $(P_{pos}$,c,o, $r_{neg})$**

- **Best for individual accuracy**

∵ Given P_{pos} , should choose r_{pos} not r_{neg}

Results

Model Variant	Accuracy ↑			Distribution ↓		
	Score	%Δ vs Base	vs GPT-4o	Score	%Δ vs Base	vs GPT-4o
Proprietary Models						
GPT-4o Base	72.9	–	–	0.174	–	–
+ Few-shot (5)	73.2	0.4%	0.4%	0.161	7.5%	7.5%
+ Reasoning	73.1	0.3%	0.3%	0.169	2.9%	2.9%
Open-Source Models						
LLaMA3-8B Base	<u>70.3</u>	–	-3.6%	0.219	–	-25.9%
+ Few-shot (5)	68.9	-2.0%	-5.5%	0.212	3.2%	-21.8%
+ Reasoning	69.8	-0.7%	-4.3%	0.174	20.6%	–
+ SFT	69.1	-1.7%	-5.2%	**0.153**	30.1%	12.1%
+ SFT w/ Reasoning	67.5	-4.0%	-7.4%	<u>0.165</u>	24.7%	5.2%
+ DPO	**72.6**	3.3%	-0.4%	0.185	15.5%	-6.3%
Qwen2.5-14B Base	<u>72.9</u>	–	–	0.205	–	-17.8%
+ Few-shot (5)	71.9	-1.4%	-1.4%	0.196	4.4%	-12.6%
+ Reasoning	72.7	-0.3%	-0.3%	0.166	19.0%	4.6%
+ SFT	69.5	-4.7%	-4.7%	**0.151**	26.3%	13.2%
+ SFT w/ Reasoning	67.6	-7.3%	-7.3%	<u>0.164</u>	20.0%	5.7%
+ DPO	**74.0**	1.4%	1.4%	0.181	11.7%	-4.0%
Bounds						
Uniform Guess	61.2	–	-16.1%	0.203	–	-16.7%
Empirical Best	–	–	–	0.125	–	28.2%

Table 2: Comparison of model variants on accuracy and distribution distance metrics across unseen studies (§5.3). In each scenario, **best scores are in boldface**, <u>second-best underlined</u>. Percent changes are relative.

Accuracy => DPO Distribution => SFT

CVS: The Flight Simulator for Management
The Problem & The Solution

Limitations of Traditional Methods

- A/B tests:

Slow (months), expensive, ethically constrained

- Surveys:

Hard-to-reach populations, fragmented findings

- Historical data:

No causal inference, no counterfactuals

Simile's Solution — Agentic Twins

- What they are:

AI agents seeded with consented, person-level data (demographics, preferences, prior behavior)

- What they can do:

Participate in surveys, navigate product experiences, interact with each other over time

Agent Creation & Validation

Agent Creation Pipeline

Step 1:

2-hour voice-to-voice interview conducted by AI interviewer, guided by sociologists at Stanford & Princeton (American Voices project)

Step 2:

Transcript -> agent memory: structured, retrievable, and updatable as new information arrives

Step 3:

Further calibration with CVS first-party data: CRM records, A/B test history, support interaction logs

Validation Results

Metric	Result
Survey outcome fidelity	~85% of human test-retest reliability
Effect-size correlation (across RCTs)	**> 0.9**
CVS internal agreement (with first-party calibration)	**up to 95%**

Four Tangible Advantages

Faster validation of known insights:

Concept-testing cycles compressed from weeks to hours; pre-screen ideas before fieldwork.

Sharper NPS and experience drivers:

Isolated which experience levers (trust in medication handling, refill ease) move satisfaction most.

Behavior change under constraints:

Tested sensitive health behaviors (adherence, shame-linked non-compliance) safely without ethical risk.

Competitive positioning:

Benchmarked vs. grocery & mass competitors; identified which investments actually move the needle.

MiroFish

Next-generation AI prediction engine powered by multi-agent technology

```
## ⚡ Overview

**MiroFish** is a next-generation AI prediction engine powered by multi-agent technology. By extracting seed information from the real world (such as breaking news, policy drafts, or financial signals), it automatically constructs a high-fidelity parallel digital world. Within this space, thousands of intelligent agents with independent personalities, long-term memory, and behavioral logic freely interact and undergo social evolution. You can inject variables dynamically from a "God's-eye view" to precisely deduce future trajectories — **rehearse the future in a digital sandbox, and win decisions after countless simulations**.

> You only need to: Upload seed materials (data analysis reports or interesting novel stories) and describe your prediction requirements in natural language</br>
> MiroFish will return: A detailed prediction report and a deeply interactive high-fidelity digital world

### Our Vision

MiroFish is dedicated to creating a swarm intelligence mirror that maps reality. By capturing the collective emergence triggered by individual interactions, we break through the limitations of traditional prediction:

- **At the Macro Level**: We are a rehearsal laboratory for decision-makers, allowing policies and public relations to be tested at zero risk
- **At the Micro Level**: We are a creative sandbox for individual users — whether deducing novel endings or exploring imaginative scenarios, everything can be fun, playful, and accessible

From serious predictions to playful simulations, we let every "what if" see its outcome, making it possible to predict anything.

## 🌐 Live Demo

Welcome to visit our online demo environment and experience a prediction simulation on trending public opinion events we've prepared for you: [mirofish-live-demo](https://666ghj.github.io/mirofish-demo/)

## 📸 Screenshots

<div align="center">
<table>
<tr>
<td><img src="./static/image/Screenshot/运行截图1.png" alt="Screenshot 1" width="100%"/></td>
<td><img src="./static/image/Screenshot/运行截图2.png" alt="Screenshot 2" width="100%"/></td>
</tr>
<tr>
<td><img src="./static/image/Screenshot/运行截图3.png" alt="Screenshot 3" width="100%"/></td>
<td><img src="./static/image/Screenshot/运行截图4.png" alt="Screenshot 4" width="100%"/></td>
</tr>
<tr>
<td><img src="./static/image/Screenshot/运行截图5.png" alt="Screenshot 5" width="100%"/></td>
<td><img src="./static/image/Screenshot/运行截图6.png" alt="Screenshot 6" width="100%"/></td>
</tr>
</table>
</div>

## 🎬 Demo Videos

### 1. Wuhan University Public Opinion Simulation + MiroFish Project Introduction
```

```
MiroFish/
├── backend/          # Flask REST API
│   └── app/
│       ├── api/      # Routes (graph, simulation, report)
│       ├── services/ # Core logic (graph builder, simulation runner, report agent)
│       ├── models/   # Project/task state management
│       └── utils/    # File parser, LLM client, logger
├── frontend/         # Vue 3 SPA
│   └── src/
│       ├── views/    # Pages (5-step workflow UI)
│       ├── components/ # Step1~5 components
│       └── api/      # Axios API client
└── locales/          # Multilingual support (EN/ZH)
```

MiroFish: Core 5-Step Workflow

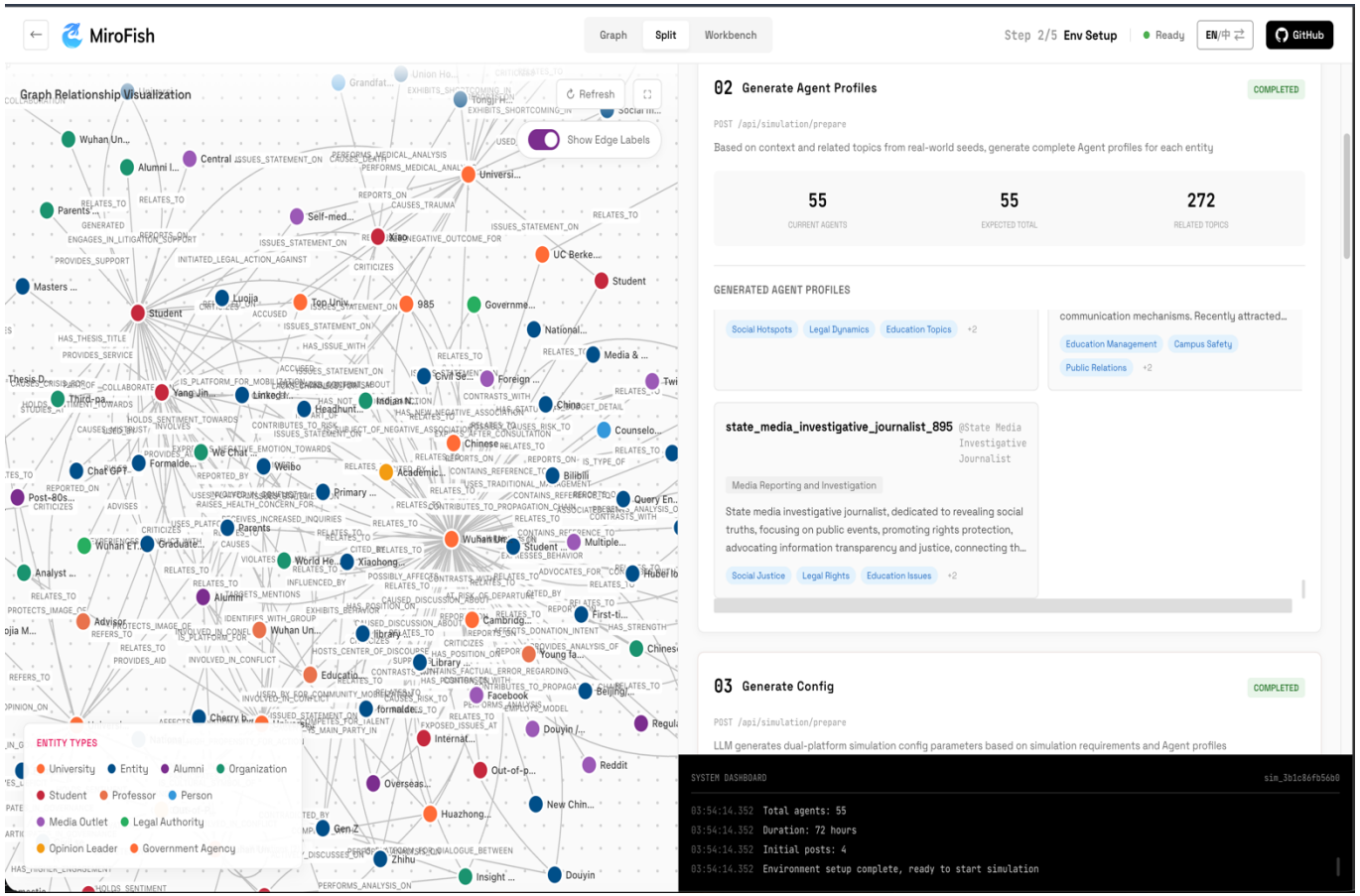
1. Graph Building:
Upload documents (PDF/TXT/MD)
→ Generate entity ontology
→ Build knowledge graph in Zep Cloud

2. Environment Setup:
Extract entities from graph → Auto-generate agent personas

3. Simulation Execution:
Run multi-agent simulation via OASIS framework (modeling Twitter/Reddit platforms)

4. Report Generation:
LLM-based ReportAgent analyzes simulation results →
Produces prediction reports

5. Deep Interaction:
Chat directly with specific agents from the simulated world



Privasis: Synthesizing the Largest “Public” Private Dataset from Scratch

Synthesis Pipeline

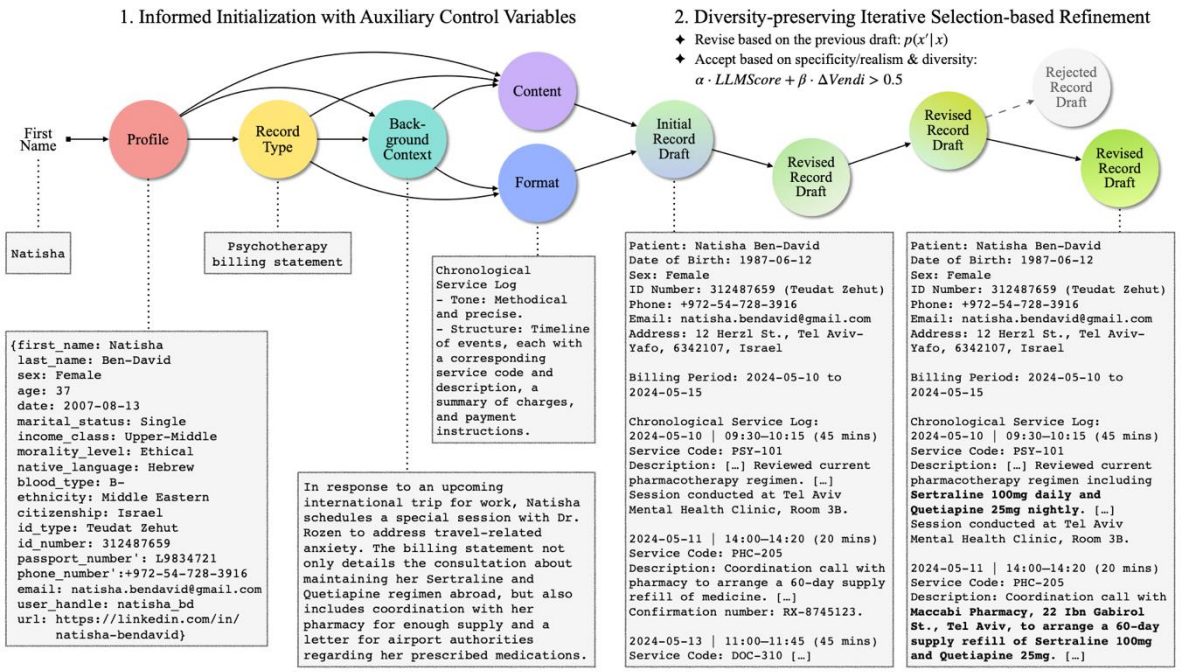


Figure 2: Overview of our synthesis pipeline.

[record x]

Patient: Natisha Ben-David
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com

2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B.

2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy,
22 Ibn Gabirol St., Tel Aviv.
60-day refill confirmed. Confirmation: RX-8745123.

2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities
regarding Sertraline and Quetiapine prescriptions.

Sanitization Pipeline – (1) Decomposition

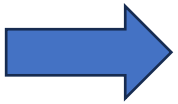
[record x]

Patient: Natisha Ben-David
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com

2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B.

2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy,
22 Ibn Gabirol St., Tel Aviv.
60-day refill confirmed. Confirmation: RX-8745123.

2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities
regarding Sertraline and Quetiapine prescriptions.



c_1:
"Patient: Natisha Ben-David
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com"

c_2:
"2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B."

c_3:
"2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy,
22 Ibn Gabirol St., Tel Aviv.
60-day refill confirmed. Confirmation: RX-8745123."

c_4:
"2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities
regarding Sertraline and Quetiapine prescriptions."

Sanitization Pipeline – (2) Target Selection

c_1:
"Patient: Natisha Ben-David
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com"

c_2:
"2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B."

c_3:
"2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy,
22 Ibn Gabirol St., Tel Aviv.
60-day refill confirmed. Confirmation: RX-8745123."

c_4:
"2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities
regarding Sertraline and Quetiapine prescriptions."

sensitivity weights:
phone_number: 0.95 ← high
email: 0.93 ← high
drug_name: 0.90 ← high
location: 0.75 ← mid (clinic + pharmacy group)
date_of_birth: 0.85 ← high
service_code: 0.20 ← low (benign)

→ Sampled targets T:
z_1 = "phone_number" → label: DROP
z_2 = "drug_name" → label: ABSTRACT
z_3 = "location" → label: ABSTRACT

Sanitization Pipeline – (3) Sanitization – (i),(ii),(iii)

sensitivity weights:
phone_number: 0.95 ← high
email: 0.93 ← high
drug_name: 0.90 ← high
location: 0.75 ← mid (clinic + pharmacy group)
date_of_birth: 0.85 ← high
service_code: 0.20 ← low (benign)

→ Sampled targets T:
z_1 = "phone_number" → label: DROP
z_2 = "drug_name" → label: ABSTRACT
z_3 = "location" → label: ABSTRACT

LLM

c_1:
"Patient: Natisha Ben-David
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com"

$instrz_1$ = "Drop the information about phone_number from the text."

c_2:
"2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B."

$instrz_2$ = "Abstract the specific drug names and dosages as 'prescribed psychiatric medications'."

$instrz_3$ = "Abstract specific clinic and pharmacy names and addresses as 'a local mental health clinic' and 'a nearby pharmacy'."

c_3:
"2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy, 22 Ibn Gabirol St., Tel Aviv.
60-day refill confirmed. Confirmation: RX-8745123."

$instrz_3$ = "Abstract specific clinic and pharmacy names and addresses as 'a local mental health clinic' and 'a nearby pharmacy'."

c_4:
"2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities regarding Sertraline and Quetiapine prescriptions."

$instrz_2$ = "Abstract the specific drug names and dosages as 'prescribed psychiatric medications'."

Sanitization Pipeline – (3) Sanitization (iv)

c_1:
"Patient: Natisha Ben-David **$instrz_1$ = "Drop …**
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com"

c_2:
"2024-05-10 | 09:30–10:15 **$instrz_2$ = "Abstract … $instrz_3$ = "Abstract …**
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B."

c_3:
"2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy,
22 Ibn Gabirol St., Tel Aviv. **$instrz_3$ = "Abstract …**
60-day refill confirmed. Confirmation: RX-8745123."

c_4:
"2024-05-13 | 11:00–11:45 **$instrz_2$ = "Abstract …**
Service Code: DOC-310
Description: Letter prepared for airport authorities
regarding Sertraline and Quetiapine prescriptions."

LLM

[sanitized record $\tilde{x}$]

Patient: Natisha Ben-David
Date of Birth: 1987-06-12
()
Email: natisha.bendavid@gmail.com

2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed prescribed psychiatric medications regimen.
Session at a local mental health clinic.

2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with a nearby pharmacy.
60-day refill confirmed. Confirmation: RX-8745123.

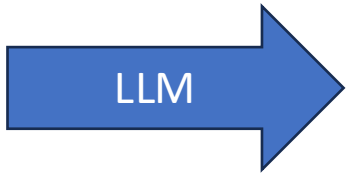
2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities
regarding prescribed psychiatric medications.

Sanitization Pipeline – (4) Final Instruction Generation

$instrz_1$ = "Drop the information about phone_number from the text."

$instrz_2$ = "Abstract the specific drug names and dosages as 'prescribed psychiatric medications'."

$instrz_3$ = "Abstract specific clinic and pharmacy names and addresses as 'a local mental health clinic' and 'a nearby pharmacy'."



Ib: "Remove the phone number, abstract specific drug names and dosages as 'prescribed psychiatric medications', and generalize clinic and pharmacy names and addresses to generic descriptions, while keeping the service codes and dates."

PRIVASIS – SANITIZATION : Dataset for Train & Benchmark

{x, lb, x̃}

[record x]

Patient: Natisha Ben-David
Date of Birth: 1987-06-12
Phone: +972-54-728-3916
Email: natisha.bendavid@gmail.com

2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed Sertraline 100mg and Quetiapine 25mg regimen.
Session at Tel Aviv Mental Health Clinic, Room 3B.

2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with Maccabi Pharmacy, 22 Ibn Gabirol St., Tel Aviv.
60-day refill confirmed. Confirmation: RX-8745123.

2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities regarding Sertraline and Quetiapine prescriptions.

**lb: "Remove the phone number, abstract specific drug names and dosages as 'prescribed psychiatric medications',
and generalize clinic and pharmacy names and addresses to generic descriptions, while keeping the service codes and dates."**

[sanitized record x̃]

Patient: Natisha Ben-David
Date of Birth: 1987-06-12
()
Email: natisha.bendavid@gmail.com

2024-05-10 | 09:30–10:15
Service Code: PSY-101
Description: Reviewed prescribed psychiatric medications regimen.
Session at a local mental health clinic.

2024-05-11 | 14:00–14:20
Service Code: PHC-205
Description: Coordination call with a nearby pharmacy.
60-day refill confirmed. Confirmation: RX-8745123.

2024-05-13 | 11:00–11:45
Service Code: DOC-310
Description: Letter prepared for airport authorities regarding prescribed psychiatric medications.

PRIVASIS – CLEANER : Sanitizer Model

37K subset of
PRIVASIS – SANITIZATION

Qwen3 models: 0.6B, 1.7B, and 4B

PRIVASIS – CLEANER

PRIVASIS – CLEANER : Result

Table 4: Sanitization performance of off-the-shelf LLMs and our PRIVASIS-CLEANER models. 'Att.' denotes attribute and 'Successful Att. / Record' denotes the average success rate within a record.

Model	Sanitization			Retention			Full
	Successful Attribute (%)	Successful Att. / Record (%)	Successful Record (%)	Successful Attribute (%)	Successful Att. / Record (%)	Successful Record (%)	Successful Record (%)
Vanilla Test Set							
o3	91.53	91.62	74.57	94.06	95.24	93.57	<u>70.25</u>
DeepSeek R1	88.54	91.01	72.26	93.74	94.39	93.38	69.58
GPT-5	90.16	91.80	73.90	93.42	94.08	92.71	70.06
GPT-4.1	88.07	90.73	72.26	94.48	95.80	94.43	70.06
GPT-OSS-120B	89.42	90.33	69.87	95.28	95.61	94.53	67.66
LLaMA-4 Maverick	89.31	90.77	73.61	89.97	90.27	88.68	67.18
Qwen3-235B	83.34	87.58	68.04	93.47	94.31	93.19	64.40
Qwen3-4B	75.66	79.59	57.58	90.66	90.95	90.02	53.65
Qwen3-0.6B	47.39	53.77	35.99	70.49	71.42	69.19	16.70
PRIVASIS-CLEANER-4B	86.60	90.60	72.84	99.15	99.13	98.66	**72.50**
PRIVASIS-CLEANER-0.6B	81.61	87.37	68.52	99.26	99.23	98.75	68.04
Hard Test Set							
o3	80.20	75.65	15.23	87.89	87.04	83.81	11.66
DeepSeek R1	75.54	72.76	15.14	86.70	86.16	82.59	11.23
GPT-5	78.78	75.30	16.28	87.08	87.23	84.25	**13.14**
GPT-4.1	75.14	72.40	13.93	89.79	90.06	86.51	12.18
GPT-OSS-120B	77.67	74.07	13.84	88.36	87.87	84.94	10.53
LLaMA-4 Maverick	76.21	73.40	16.19	82.07	81.92	78.24	11.05
Qwen3-235B	69.01	67.33	12.79	89.37	89.71	86.95	10.27
Qwen3-4B	62.15	59.89	8.27	84.36	84.01	80.68	5.66
Qwen3-0.6B	33.41	35.46	12.53	76.63	76.08	72.58	4.44
PRIVASIS-CLEANER-4B	73.29	71.23	14.19	95.76	95.60	92.96	<u>12.36</u>
PRIVASIS-CLEANER-0.6B	68.11	66.14	11.31	95.09	95.20	91.99	9.31

PRIVASIS + MemAgent => PrivasisAgent

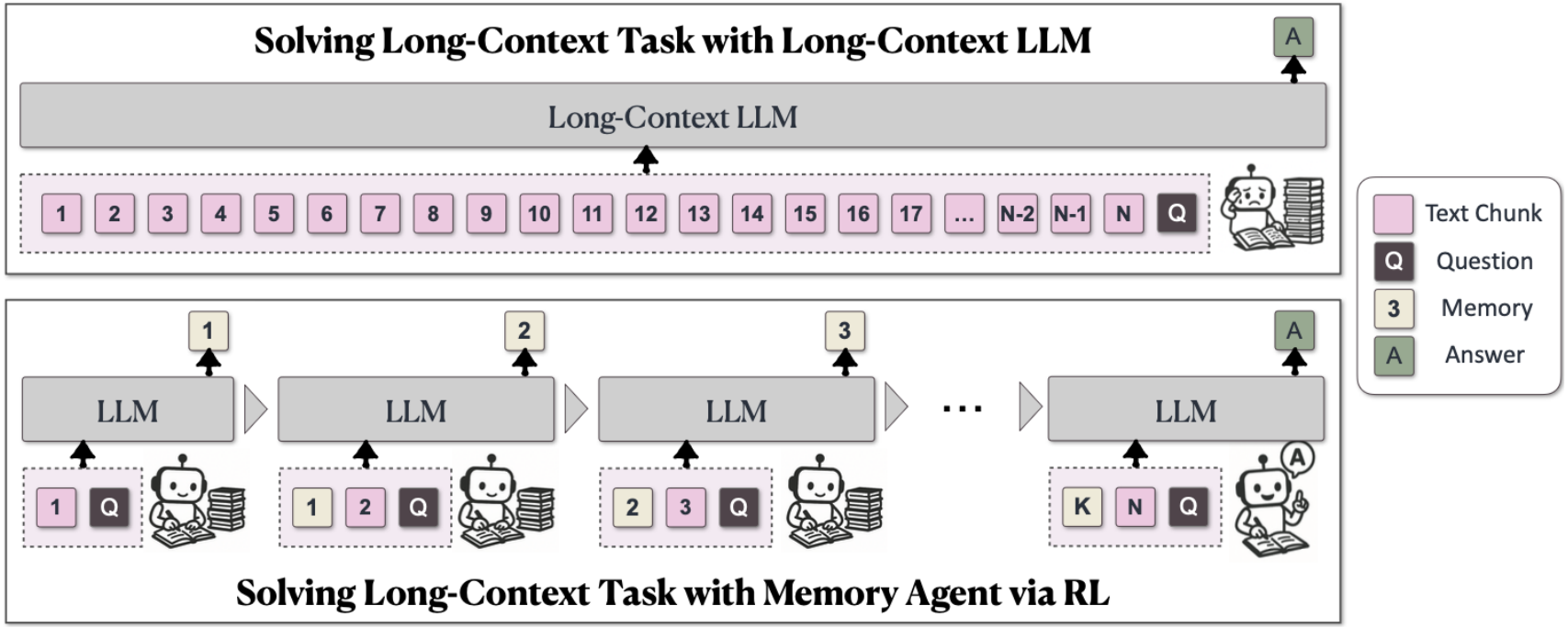


Figure 2 MemAgent is inspired by the way humans process long documents. It divides the document into multiple chunks and allows LLMs to process them iteratively, recording relevant information in memory. Finally, LLMs generate answers based on the information stored in the memory.

Replace Text Chunk => Privacy Chunk

Agent: Sanitizing prompt + RL