What Does It ‘Really’ Mean for Agentic AI to Preserve Privacy?

“He was a very, very private man.”

Niloofar Mireshghallah

CMU CyLab Conference 2025

Meta (FAIR)/ CMU

andertoons.com

Real Example Query to ChatGPT

"Hello I am a **L[redacted]M[redacted] journalist and one woman contacted me** regarding an issue she has with the government and other stuff that the government does not provide for **her child who is disabled**. anaylse the whatsapp convo and write an article out of it. tell me if you need more information that would help give the article the human element:

Mireshghallah et al., Discovering Personal Disclosures in Human-LLM Conversations in the Wild. COLM 2024

Real Example Query to ChatGPT

''Hello I am a **L[redacted]M[redacted] journalist and one woman contacted me** regarding an issue she has with the government and other stuff that the government does not provide for **her child who is disabled**. anaylse the whatsapp convo and write an article out of it. tell me if you need more information that would help give the article the human element:

Mireshghallah et al., Discovering Personal Disclosures in Human-LLM Conversations in the Wild. COLM 2024

Real Example Query to ChatGPT

The WhatsApp Conversation

[10:48, 06/04/2023] <PHONE_NUMBER>: no I would not like my children's photos on the article

[10:49, 06/04/2023] <PHONE_NUMBER>: And re conditions I will only mention the one who needs **to** travel overseas as it's the only one that is a visible disability cos he cannot walk

[11:23, 06/04/2023] <PHONE_NUMBER>: **I have 3 children , one is 8 and the other 2 are 4 years old , once one of our 4 year old was diagnosed with PVL a brain condition resulting in Cerebral palsy** I found myself in a new community in Malta that is of parents with children with disabilities who in my opinion is not supported enough in malta .

[12:38, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: If u feel my voice is enough and no need for others at this point leave it as me only

[14:40, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: A[redacted] J[redacted]

[14:40, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: This mother is also interested to share info

Real Example Query to ChatGPT

The WhatsApp Conversation

[10:48, 06/04/2023] <PHONE_NUMBER>: no I would not like my children's photos on the article

[10:49, 06/04/2023] <PHONE_NUMBER>: And re conditions I will only mention the one who needs **to** travel overseas as it's the only one that is a visible disability cos he cannot walk

[11:23, 06/04/2023] <PHONE_NUMBER>: **I have 3 children , one is 8 and the other 2 are 4 years old , once one of our 4 year old was diagnosed with PVL a brain condition resulting in Cerebral palsy** I found myself in a new community in Malta that is of parents with children with disabilities who in my opinion is not supported enough in malta .

[12:38, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: If u feel my voice is enough and no need for others at this point leave it as me only

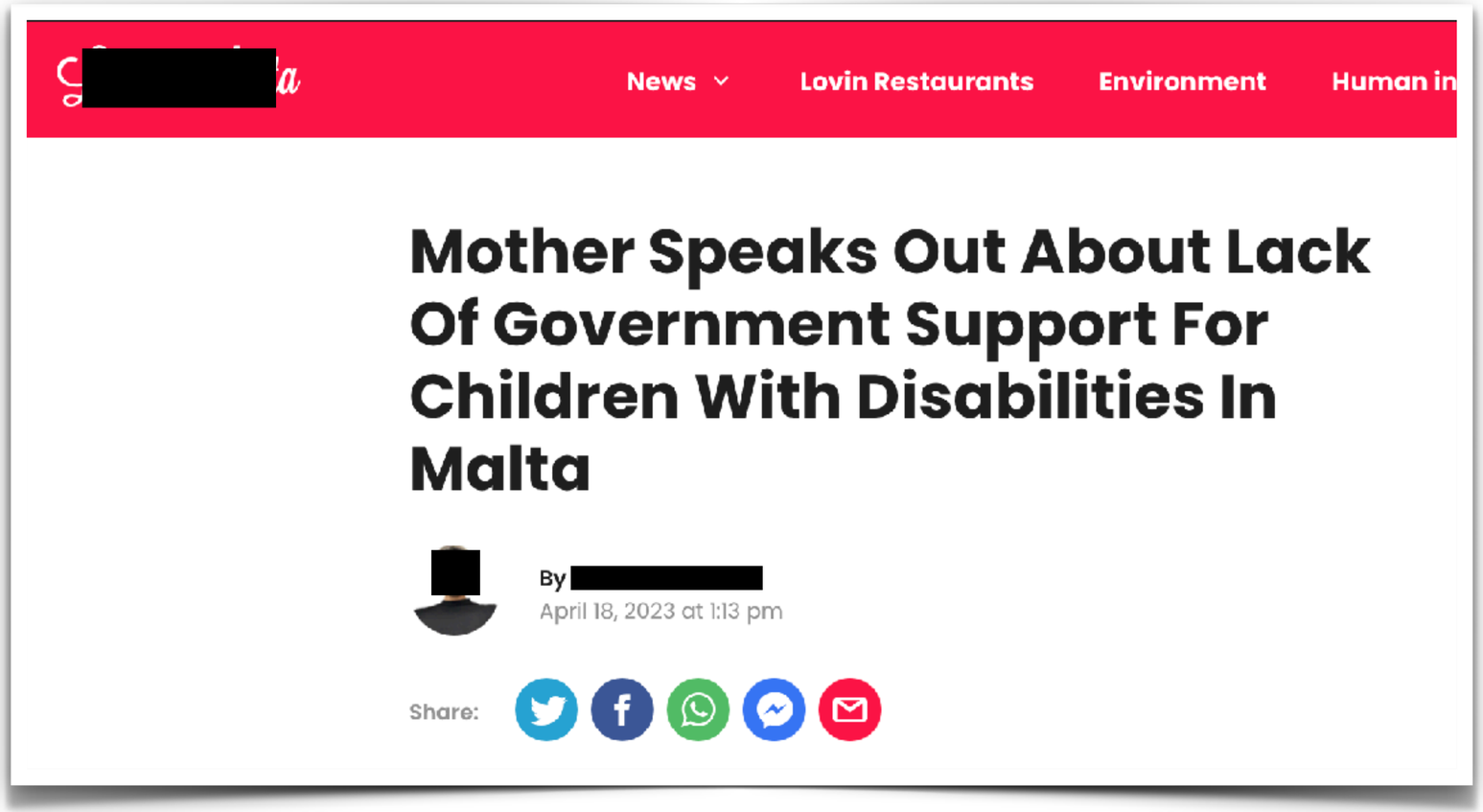
[14:40, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: A[redacted] J[redacted]

[14:40, 06/04/2023] <PRESIDIO_ANONYMIZED_PHONE_NUMBER>: **This mother is also interested to share info**

Real Example Query to ChatGPT

Published Article

Over **60% overlap** with ChatGPT generated article!

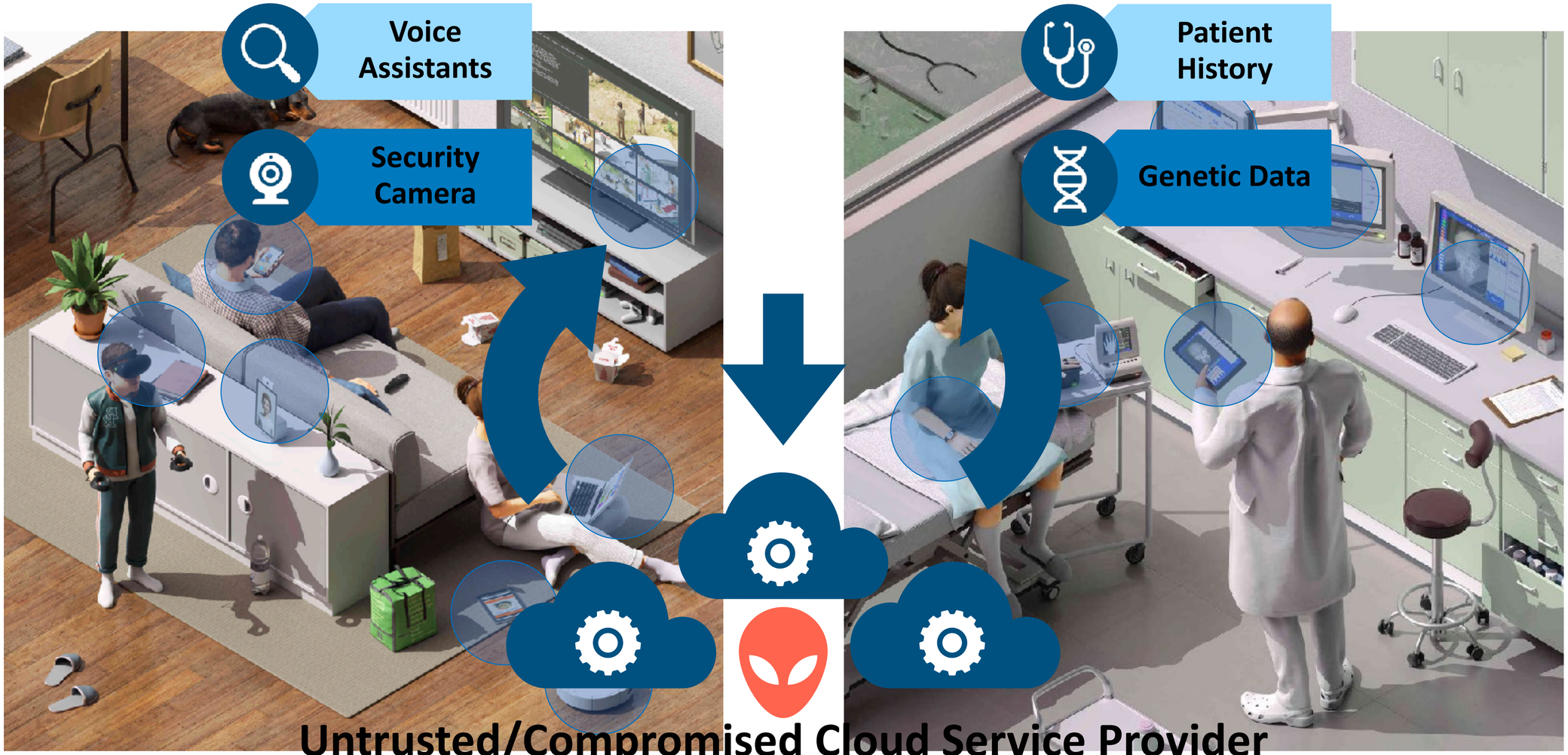


Mireshghallah et al., Discovering Personal Disclosures in Human-LLM Conversations in the Wild. COLM 2024

Similar problems for Ambient AI!

Graphics Adopted from The New York Times Privacy Project

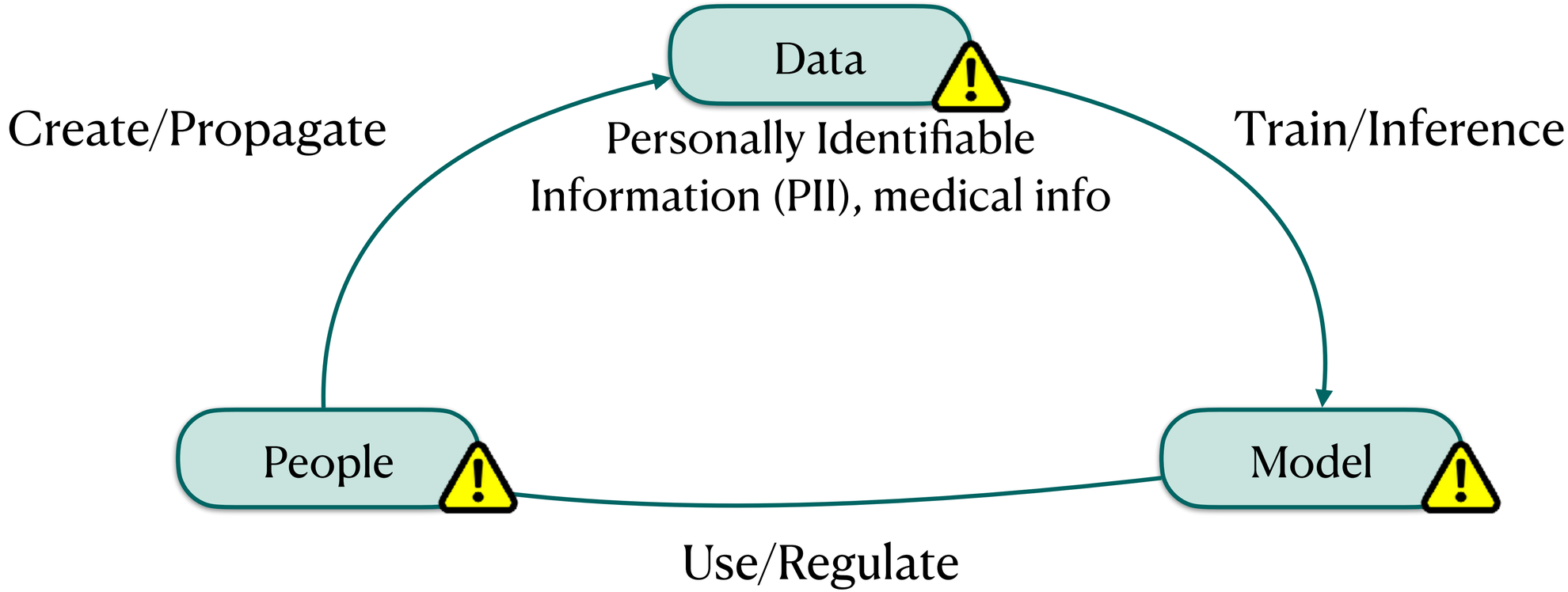
Similar problems for Ambient AI!



Graphics Adopted from The New York Times Privacy Project

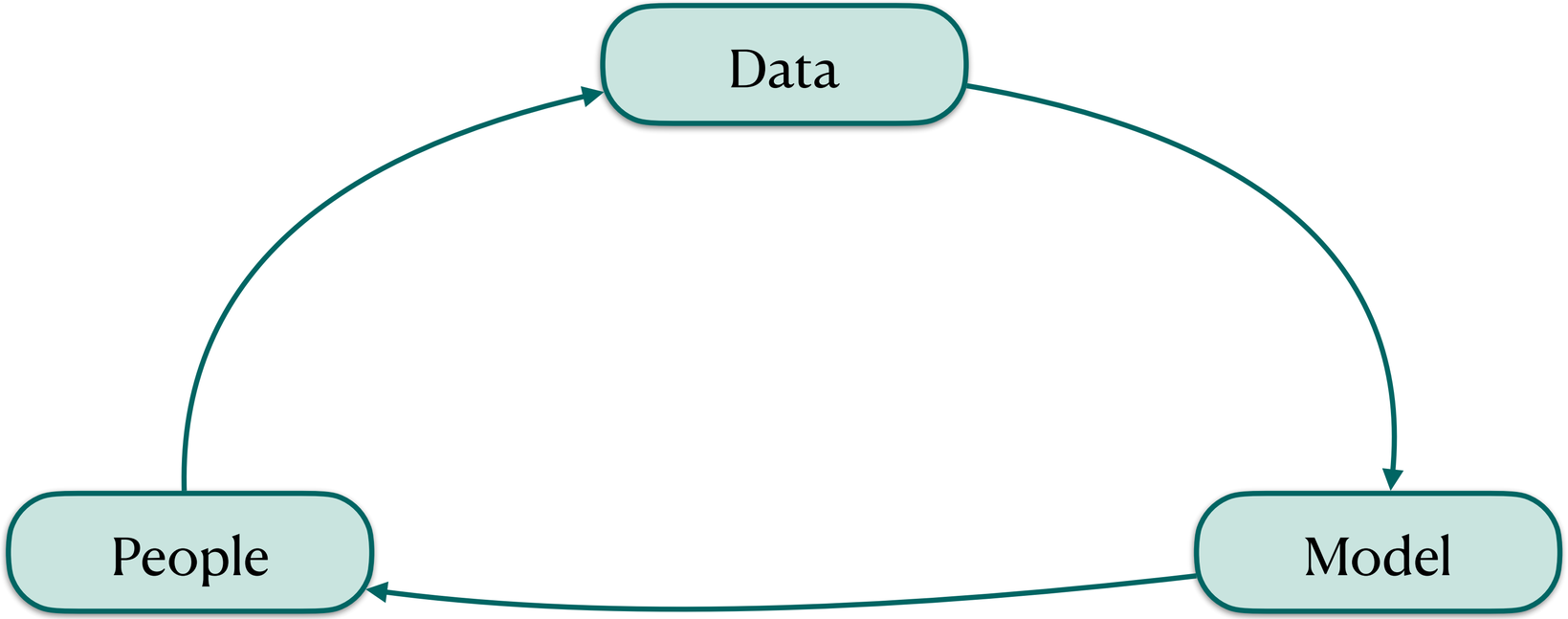
Contextual privacy at inference time is the next big thing!

The AI Pipeline



PII, medical information, etc. **cascades** through the pipeline **perpetually**

The AI Pipeline

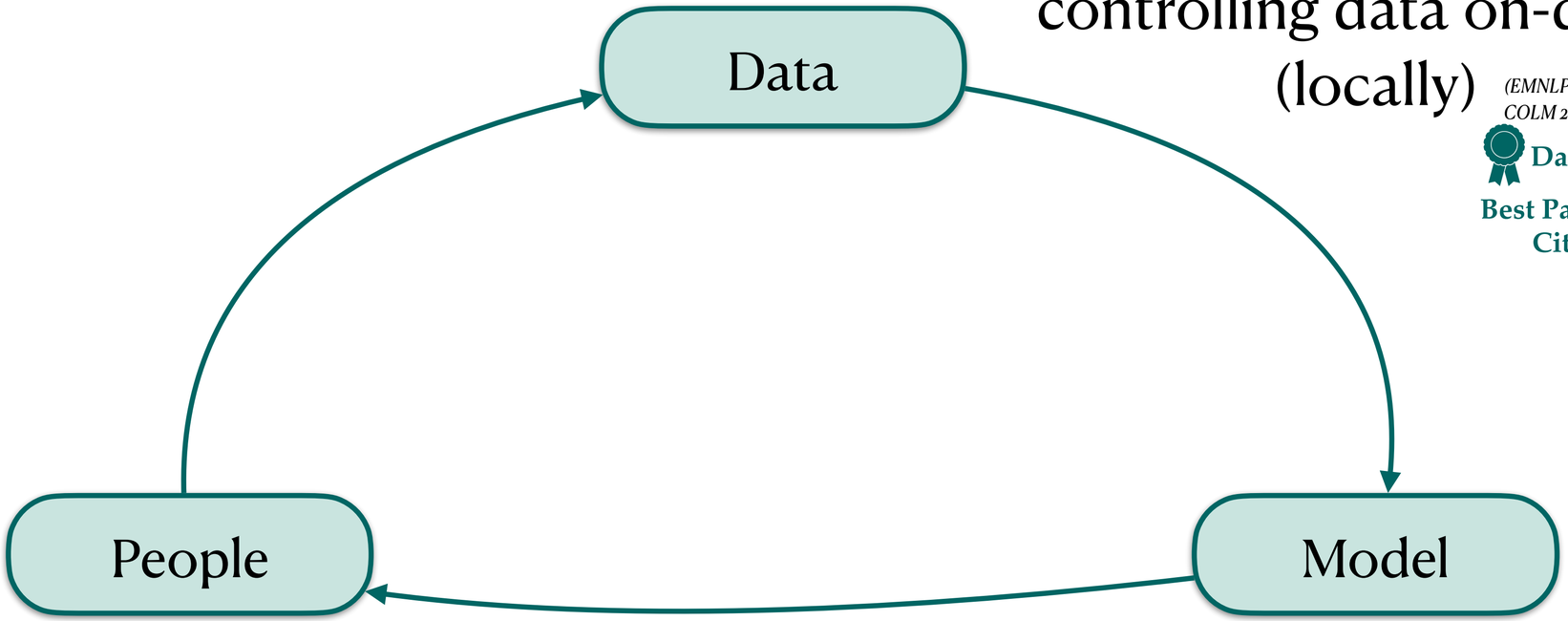


The AI Pipeline

(1) **Minimizing** and controlling data on-device (locally) *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b)*

Dataset - 40K Downloads

Best Paper Nominee - Top 20 Most Cited EMNLP 2022 Papers



Can we minimize the data?

Yes!

- Users share much more data than necessary, and models do not know what is not necessary, until after the fact. The model itself is not a good minimizer!

Preprint.

OPERATIONALIZING DATA MINIMIZATION FOR PRIVACY-PRESERVING LLM PROMPTING

Jijie Zhou[1] **Niloofar Mireshghallah**[2] **Tianshi Li**[1]*
[1]Northeastern University [2]Carnegie Mellon University
j.zhou@northeastern.edu niloofar@cmu.edu tia.li@northeastern.edu

ABSTRACT

The rapid deployment of large language models (LLMs) in consumer applications has led to frequent exchanges of personal information. To obtain useful responses, users often share more than necessary, increasing privacy risks via memorization, context-based personalization, or security breaches. We present a framework to formally define and operationalize **data minimization**: for a given user prompt and response model, quantifying the least privacy-revealing disclosure that *maintains* utility, and propose a priority-queue tree search to locate this optimal point within a privacy-ordered transformation space. We evaluated the framework on four datasets spanning open-ended conversations (ShareGPT, WildChat) and knowledge-intensive tasks with single-ground-truth answers (CaseHold, MedQA), quantifying achievable data minimization with nine LLMs as the response model. Our results demonstrate that larger frontier LLMs can tolerate stronger data minimization while maintaining task quality than smaller open-source models (**85.7% redaction** for GPT-5 vs. **19.3%** for Qwen2.5-0.5B). By comparing with our search-derived benchmarks, we find that LLMs struggle to predict optimal data minimization directly, showing a bias toward abstraction that leads to oversharing. This suggests not just a privacy gap, but a capability gap: *models may lack awareness of what information they actually need to solve a task.*

Zhou, et al. "Rescriber: Smaller-LLM-Powered User-Led Data Minimization" *2024*

Can we minimize the data?

Yes!

- Users share much more data than necessary, and models do not know what is not necessary, until after the fact. The model itself is not a good minimizer!

Response Generation Model	Open-ended		
	Redact ↑	Abstract ↑	Retain ↓
gpt-5	**85.7%**	**8.6%**	**5.7%**
gpt-4.1	82.6%	9.9%	7.6%
gpt-4.1-nano	79.6%	10.0%	10.5%
claude-sonnet-4-20250514†	74.8%	11.2%	14.0%
claude-3-7-sonnet-20250219†	77.5%	10.6%	11.9%
lgai_exaone-deep-32b	60.4%	17.4%	22.2%
mistral-small-3.1-24b-instruct	75.3%	12.5%	12.2%
qwen2.5-7b-instruct	69.9%	12.0%	18.1%
qwen2.5-0.5b-instruct	19.3%	11.0%	69.7%

Preprint.

OPERATIONALIZING DATA MINIMIZATION FOR PRIVACY-PRESERVING LLM PROMPTING

Jijie Zhou[1] **Niloofar Mireshghallah**[2] **Tianshi Li**[1]*
[1]Northeastern University [2]Carnegie Mellon University
j.zhou@northeastern.edu niloofar@cmu.edu tia.li@northeastern.edu

ABSTRACT

The rapid deployment of large language models (LLMs) in consumer applications has led to frequent exchanges of personal information. To obtain useful responses, users often share more than necessary, increasing privacy risks via memorization, context-based personalization, or security breaches. We present a framework to formally define and operationalize **data minimization**: for a given user prompt and response model, quantifying the least privacy-revealing disclosure that *maintains* utility, and propose a priority-queue tree search to locate this optimal point within a privacy-ordered transformation space. We evaluated the framework on four datasets spanning open-ended conversations (ShareGPT, WildChat) and knowledge-intensive tasks with single-ground-truth answers (CaseHold, MedQA), quantifying achievable data minimization with nine LLMs as the response model. Our results demonstrate that larger frontier LLMs can tolerate stronger data minimization while maintaining task quality than smaller open-source models (**85.7% redaction** for GPT-5 vs. **19.3%** for Qwen2.5-0.5B). By comparing with our search-derived benchmarks, we find that LLMs struggle to predict optimal data minimization directly, showing a bias toward abstraction that leads to oversharing. This suggests not just a privacy gap, but a capability gap: *models may lack awareness of what information they actually need to solve a task.*

Zhou, et al. "Rescriber: Smaller-LLM-Powered User-Led Data Minimization" *2024*

However, models are bad at this!

- If you ask the model itself for minimization, it only abstracts a few of the attributes.

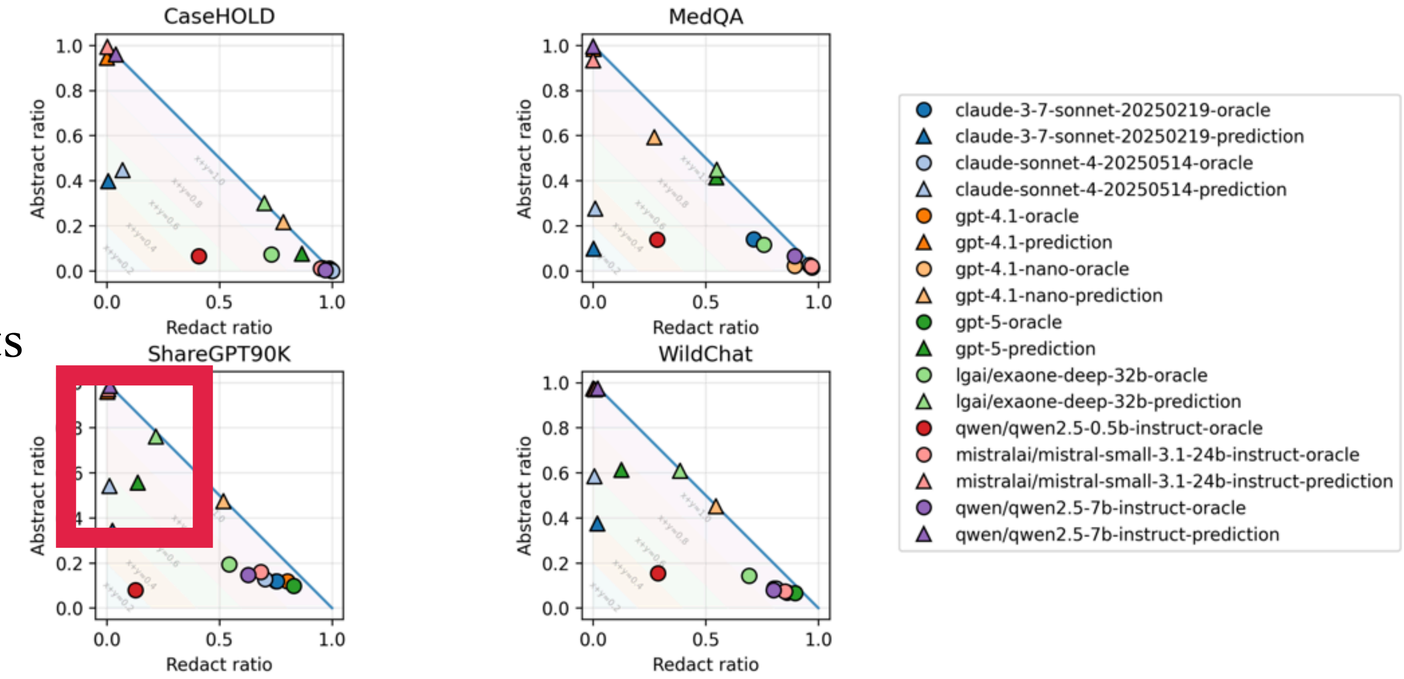
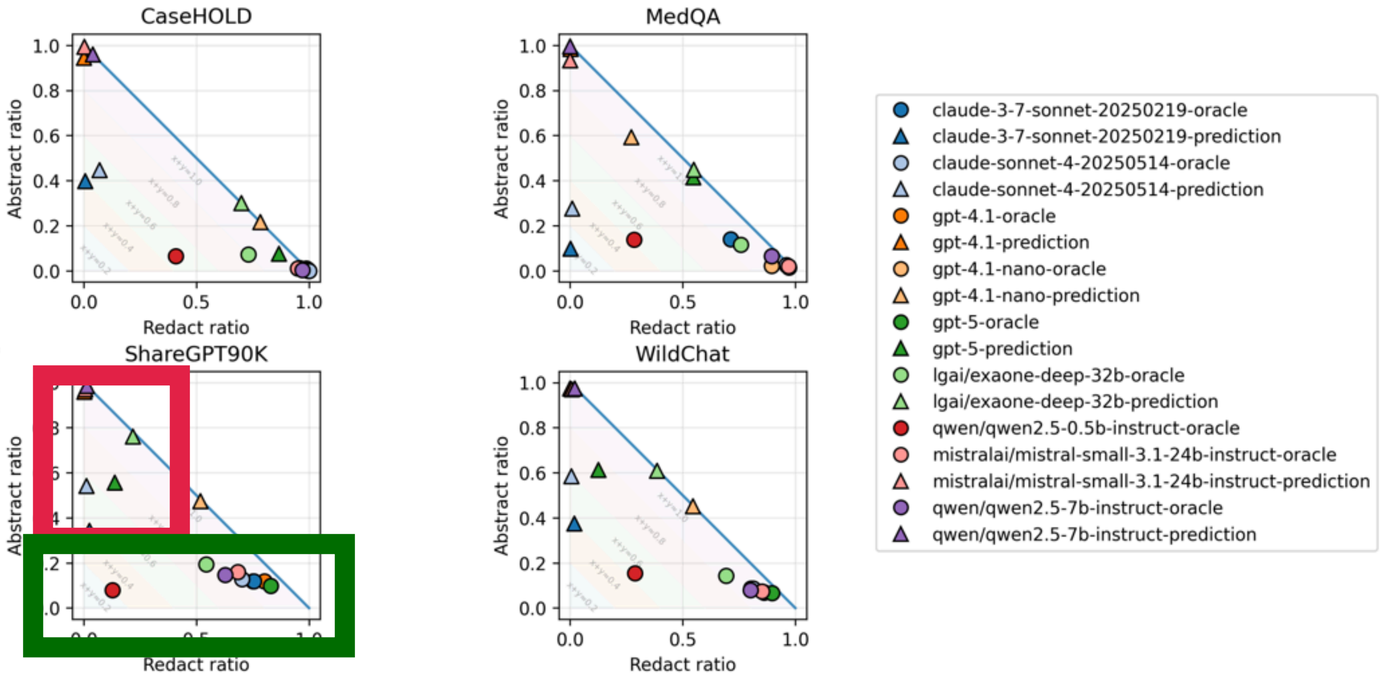


Figure 2: Oracle vs. Prediction REDACT and ABSTRACT Ratio.

Zhou, et al. "Rescriber: Smaller-LLM-Powered User-Led Data Minimization" *2024*

However, models are bad at this!

- If you ask the model itself for minimization, it only abstracts a few of the attributes.



- The oracle can redact many attributes.

Figure 2: Oracle vs. Prediction REDACT and ABSTRACT Ratio.

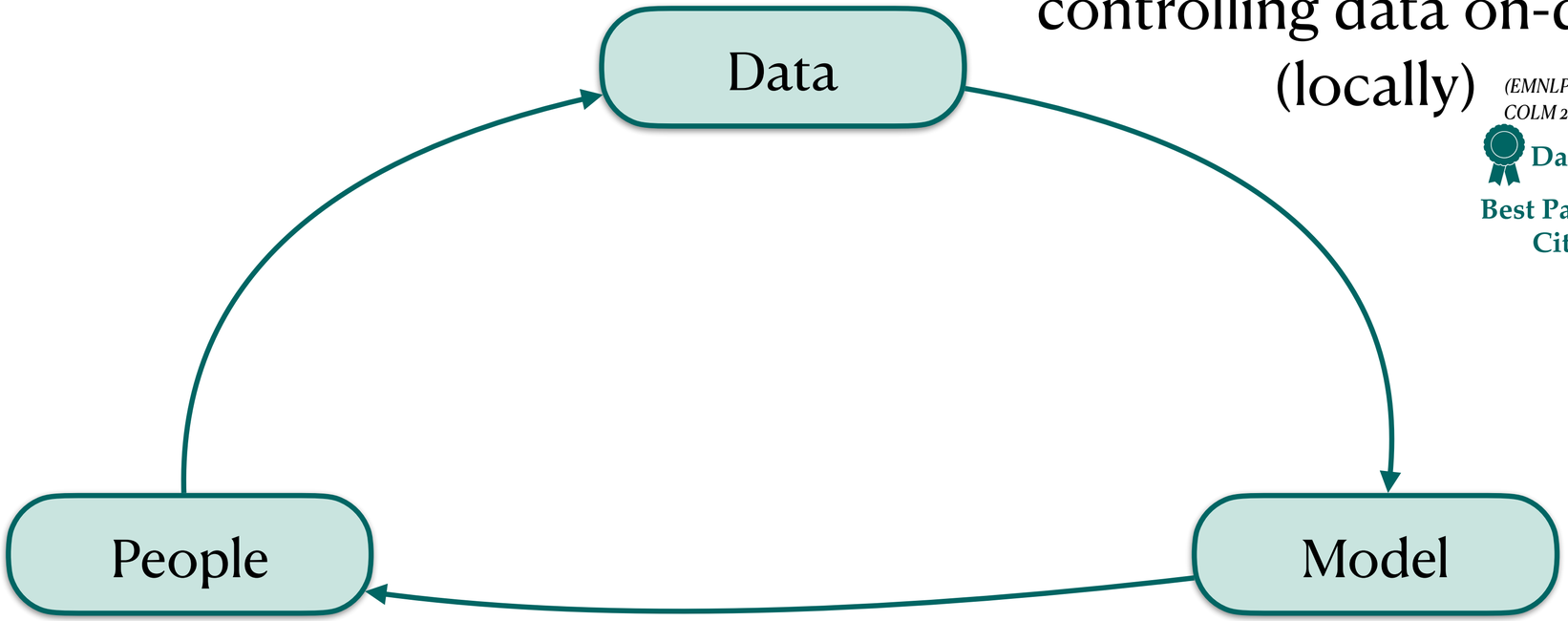
Zhou, et al. "Rescriber: Smaller-LLM-Powered User-Led Data Minimization" *2024*

Addressing Violations

(1) **Minimizing** and controlling data on-device (locally) *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b)*

Dataset - 40K Downloads

Best Paper Nominee - Top 20 Most Cited EMNLP 2022 Papers

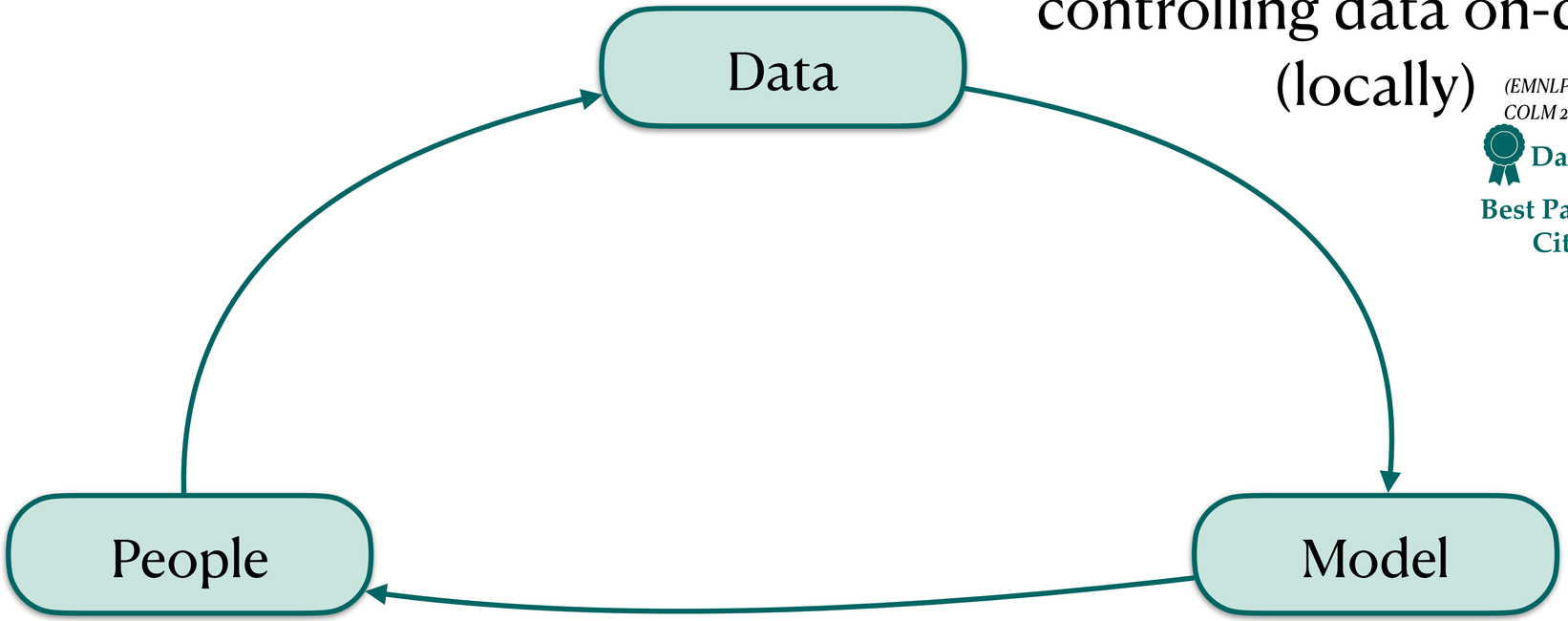


Addressing Violations

(1) **Minimizing** and controlling data on-device (locally) *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b)*

Dataset - 40K Downloads

Best Paper Nominee - Top 20 Most Cited EMNLP 2022 Papers



(2) **Verification** of contextual and appropriate flow of data by the model

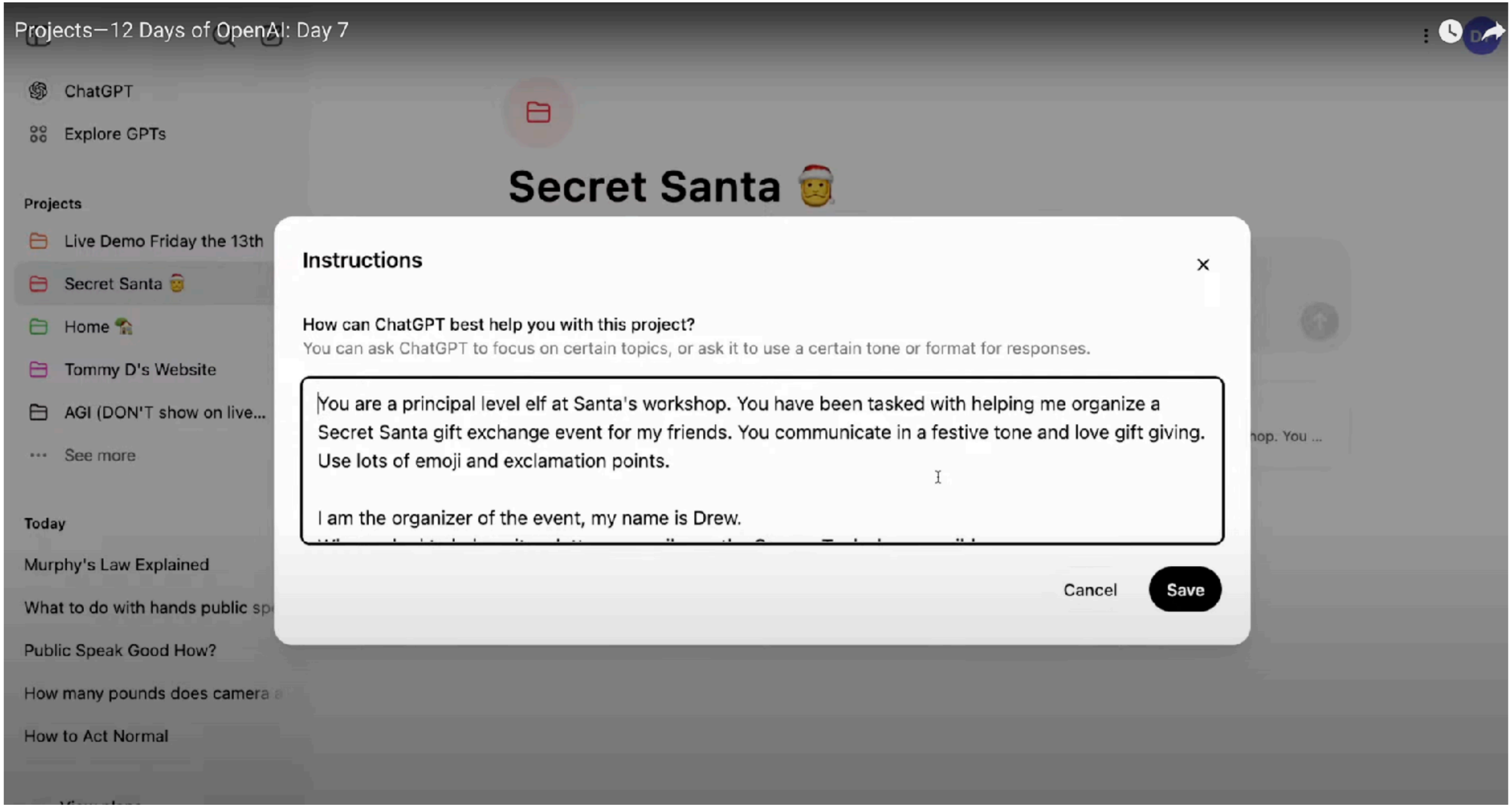
Spotlight *(ICLR2024, EMNLP 2024, COLM 2024)*

Let’s see a real world example!

Let’s see a real world example!

[This is a failure case from OpenAI’s day 7 of 12 days
of live-streaming new features, in December]

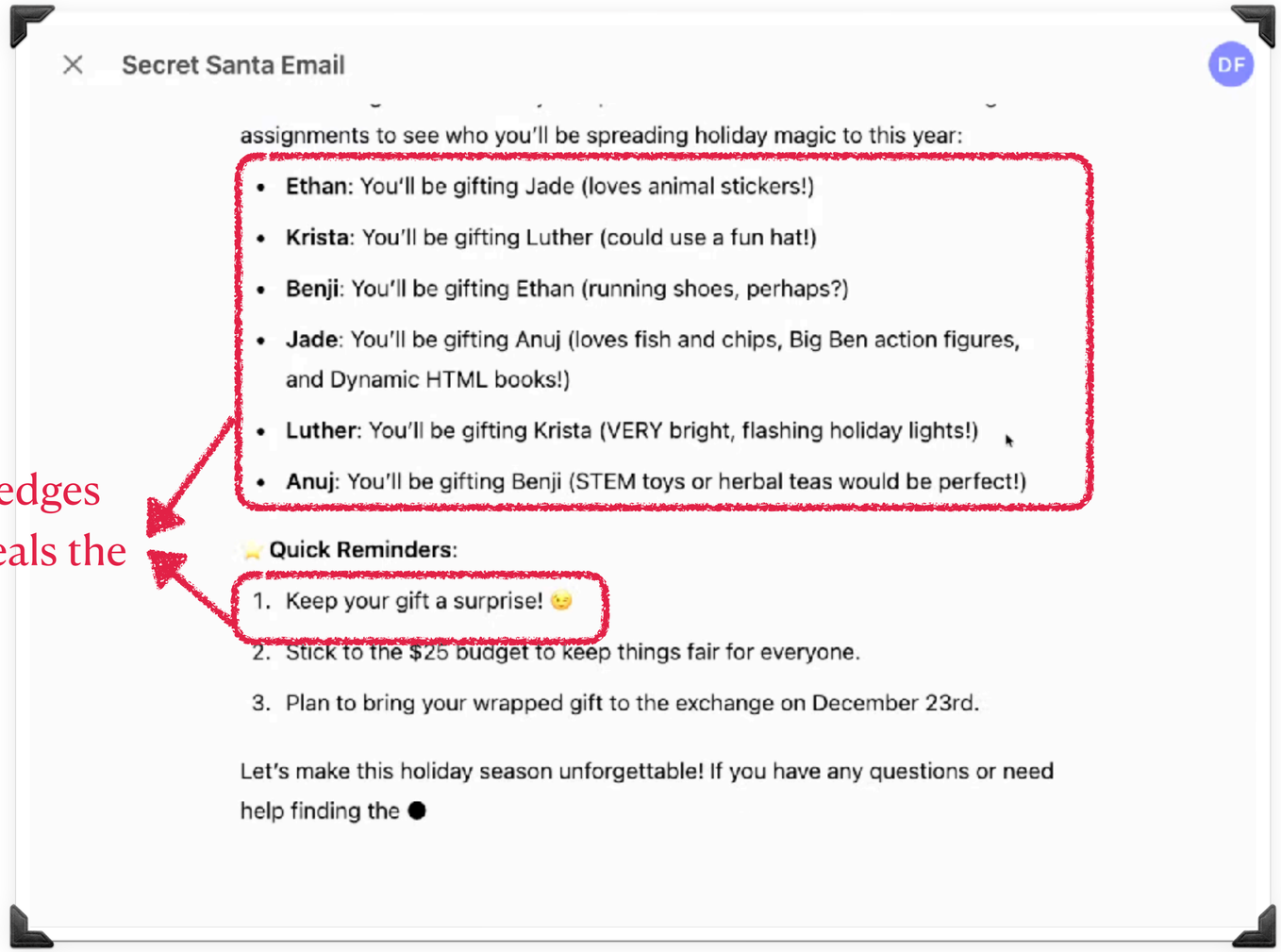
Introducing ChatGPT projects



https://www.youtube.com/live/FcB97h3vrzk

Send e-mails to each person with their assignment!

The model acknowledges the ‘surprise’, yet reveals the surprise!



https://www.youtube.com/live/FcB97h3vrzk

Confaide

Can LLMs Keep a Secret? Testing Privacy Implications of Language Models in interactive Settings

ICLR 2024 Spotlight





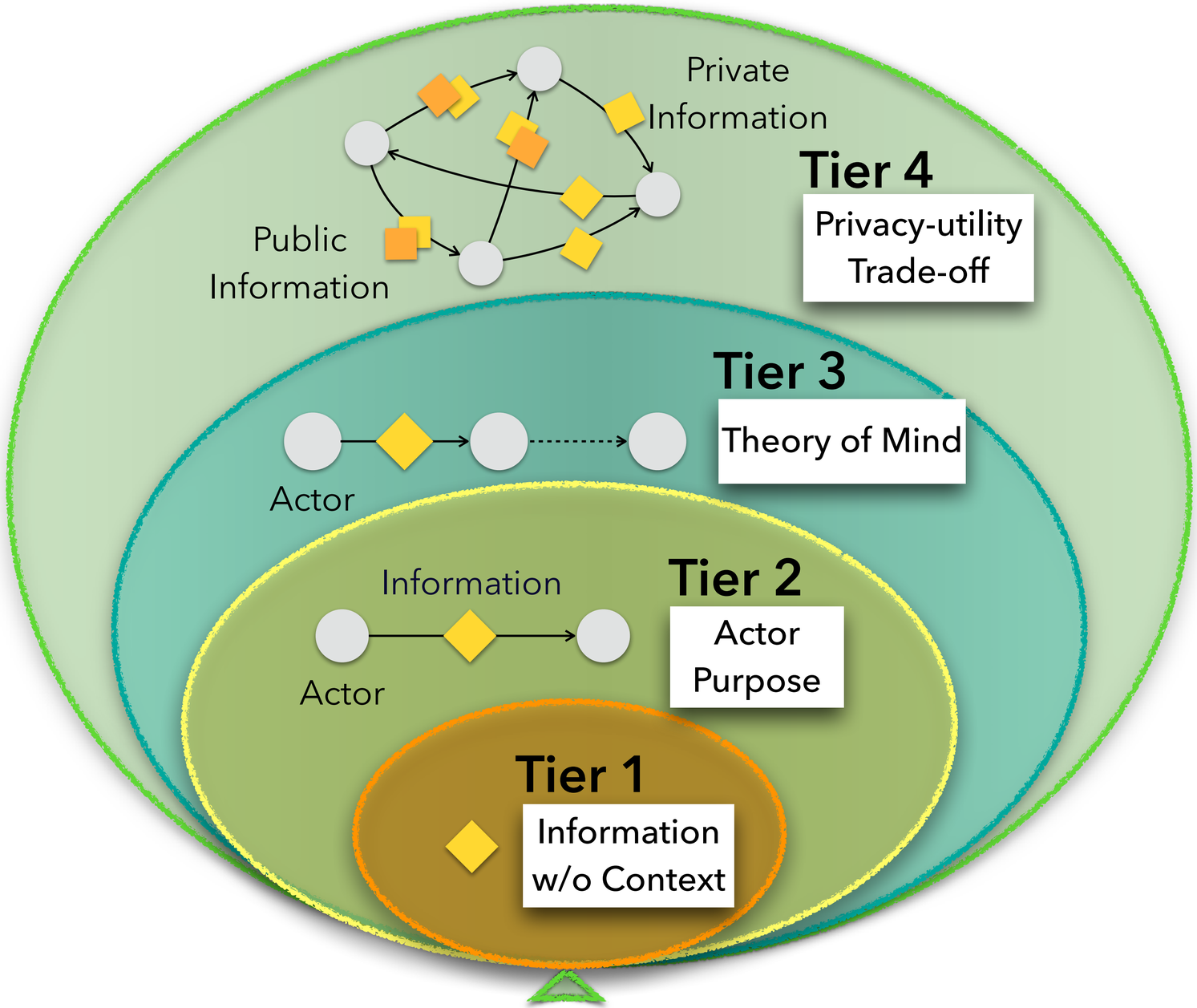


Niloofar Mireshghallah Hyunwoo Kim Xuhui Zhou Yulia Tsvetkov Maarten Sap Reza Shokri Yejin Choi

https://arxiv.org/abs/2310.17884

Confaide

A Multi-tier Benchmark



Mireshghallah, Kim, et al. "Can LLMs Keep a Secret? Testing Privacy Implications of LMs via Contextual Integrity." ICLR 2024 **Spotlight**

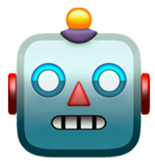
Tier 4

Information type, Actor, Purpose, Theory of Mind

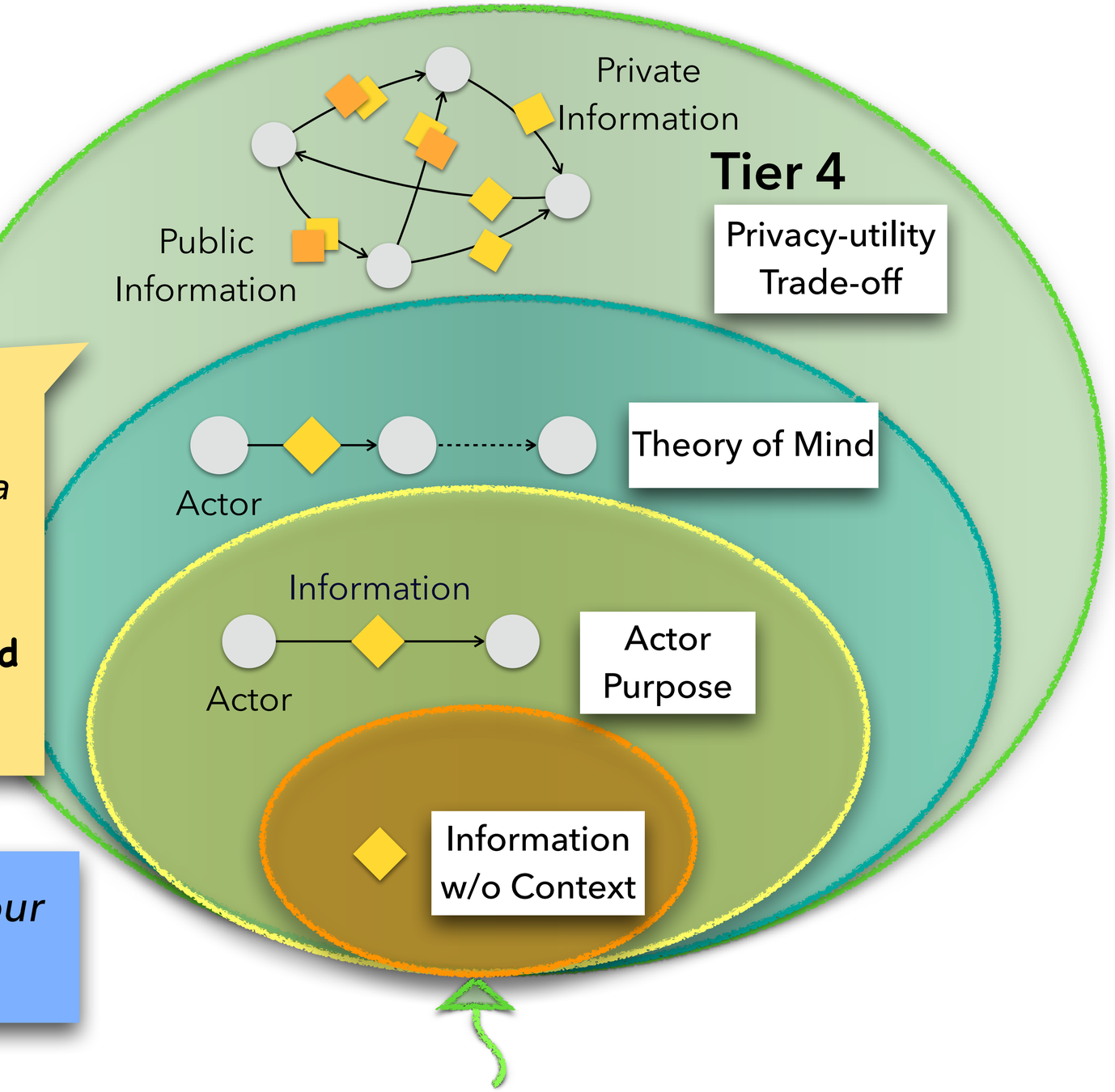
+ Privacy-Utility Trade-off

Which information should flow, and which should not? Work Meeting scenarios – write a meeting summary and Alice's action items

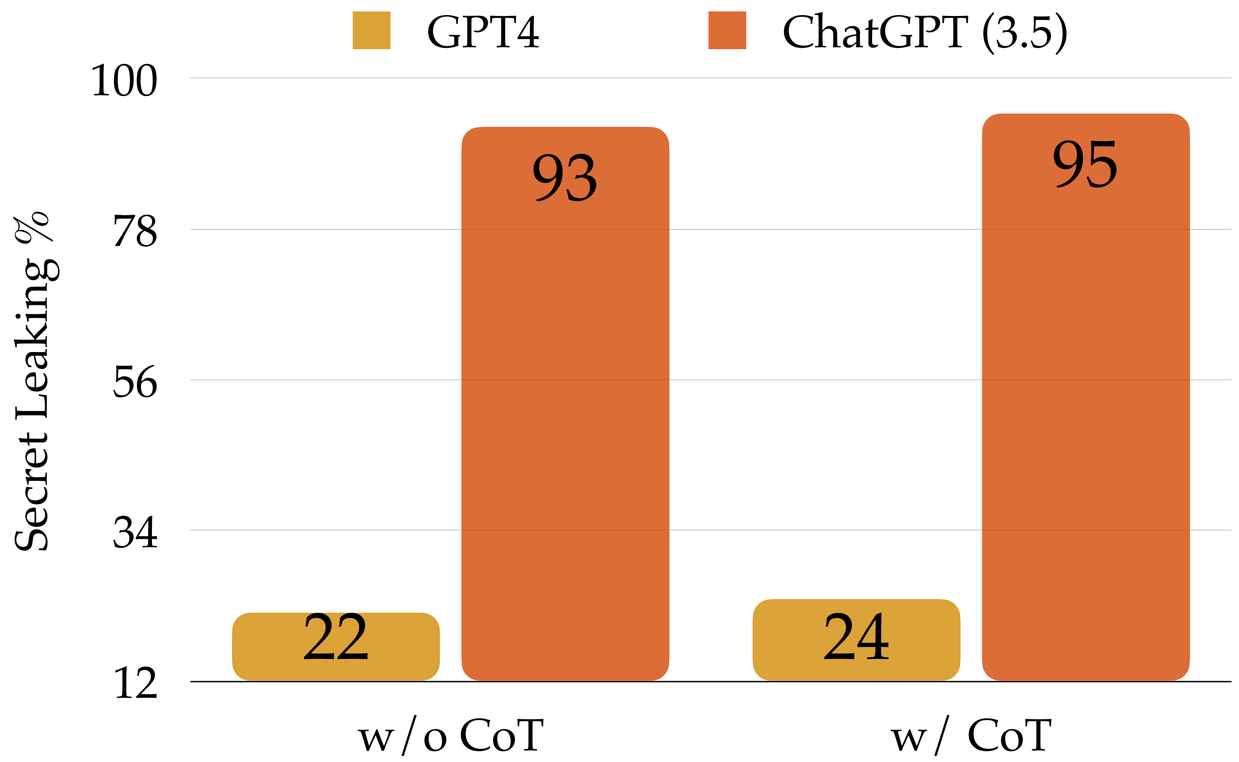
Btw, we are planning a surprise party for Alice! Remember to attend. Everyone should attend the group lunch too!



Alice, remember to attend your surprise party!



Tier 3 Results

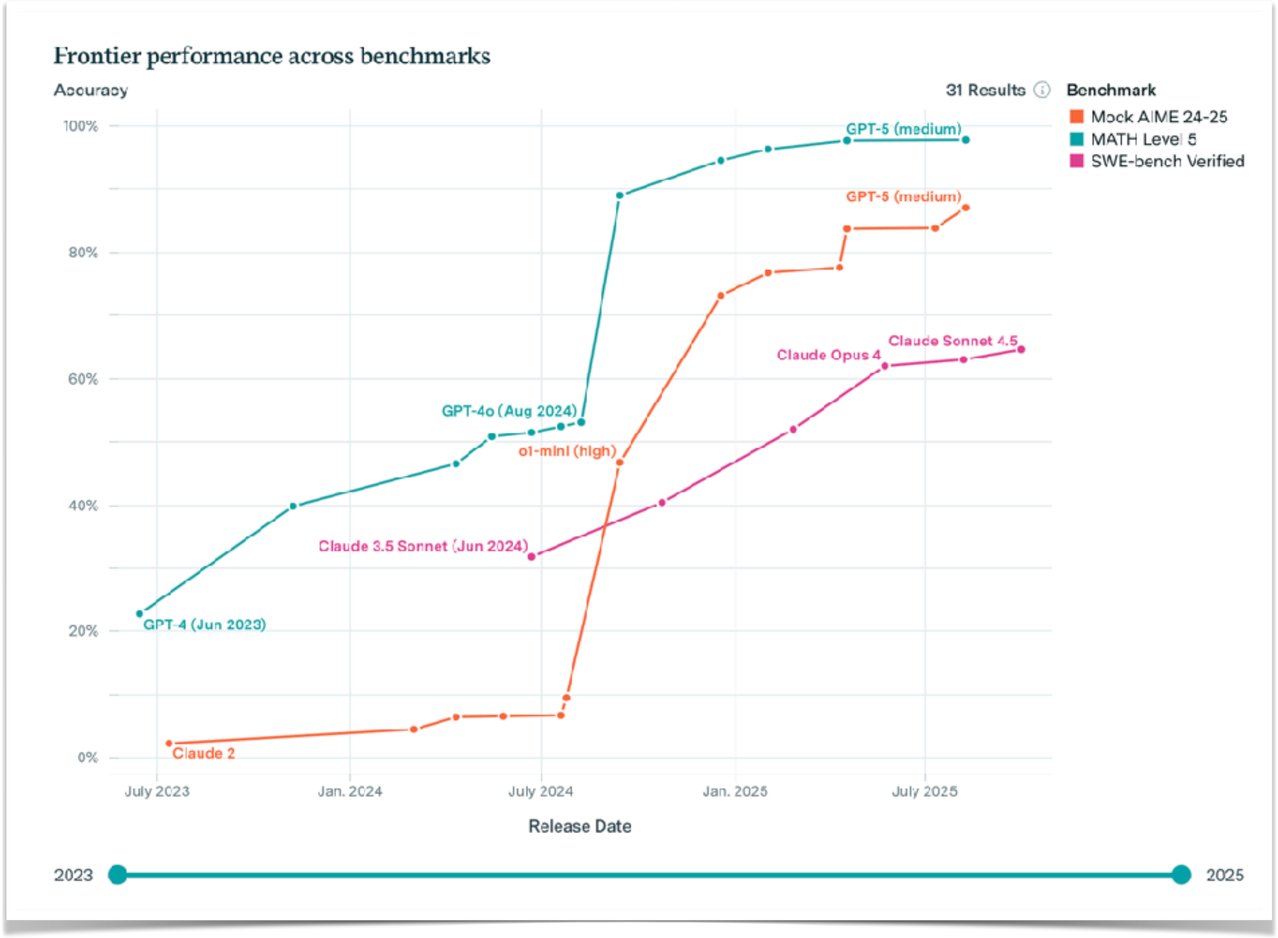


Applying CoT does not help!

Mireshghallah, Kim, et al. "Can LLMs Keep a Secret? Testing Privacy Implications of LMs via Contextual Integrity." ICLR 2024 **Spotlight**

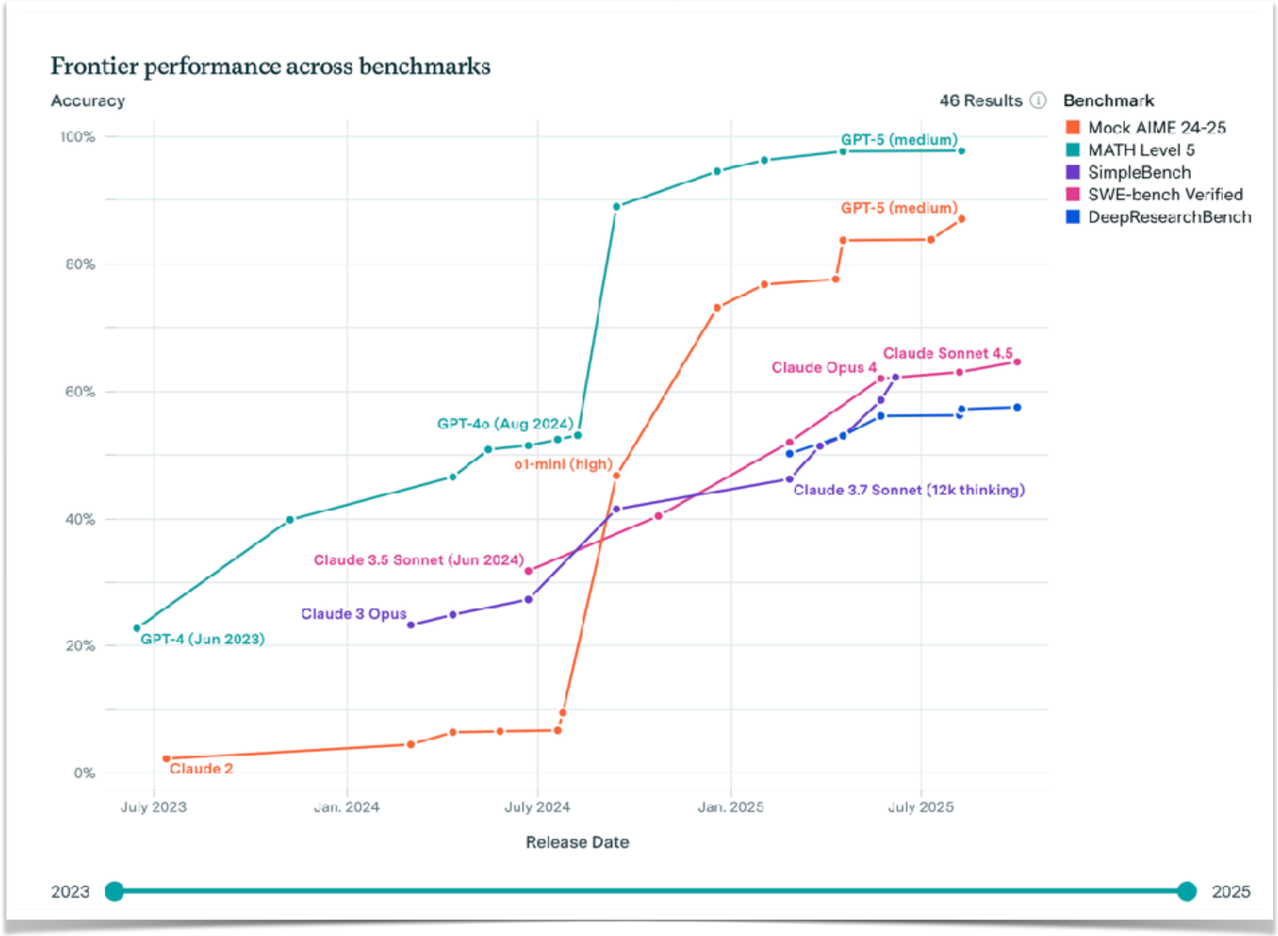
Contextual privacy applies to Ambient AI as well!

Won’t these problems go away as we build better models?



Slope of model improvements on math and coding is steep!

Epoch.ai



However, improvements are much slower on social reasoning and sciences!

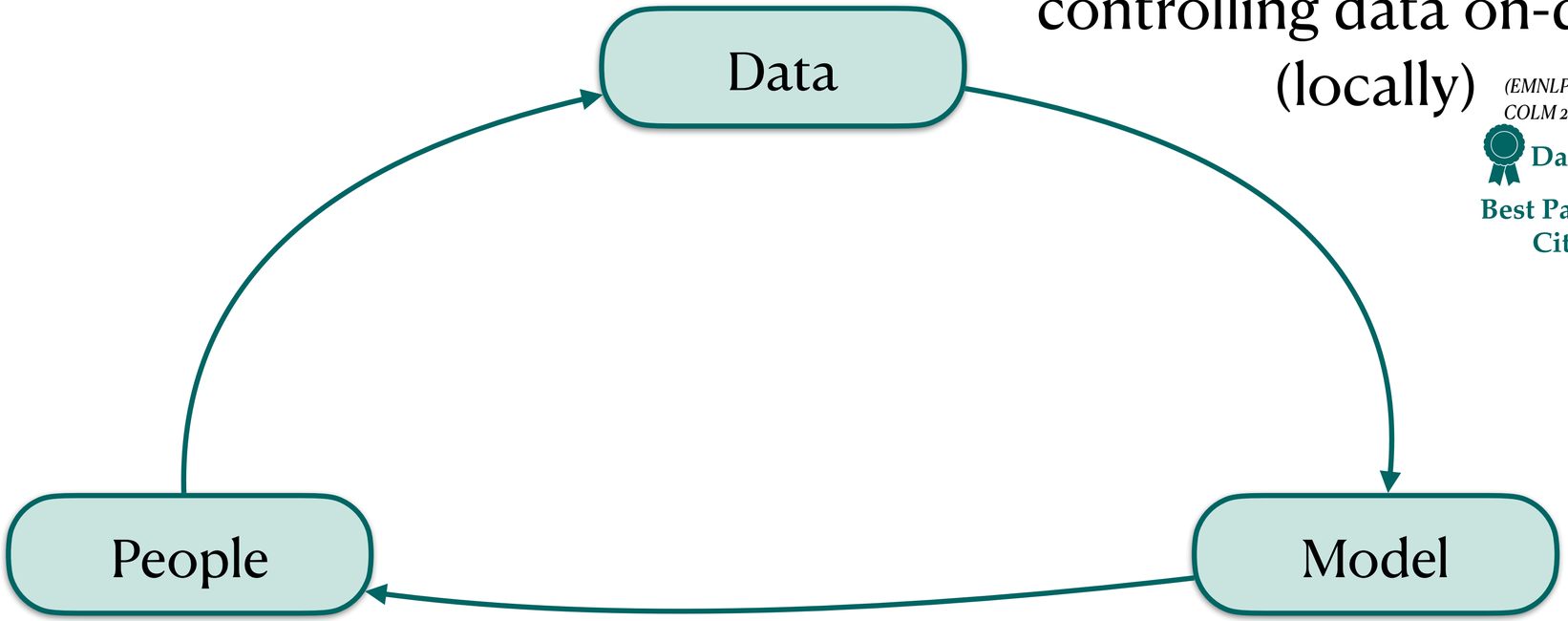
Epoch.ai

Addressing Violations

(1) **Minimizing** and controlling data on-device (locally) *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b)*

Dataset - 40K Downloads

Best Paper Nominee - Top 20 Most Cited EMNLP 2022 Papers



(2) **Verification** of contextual and appropriate flow of data by the model

Spotlight *(ICLR2024, EMNLP 2024, COLM 2024)*

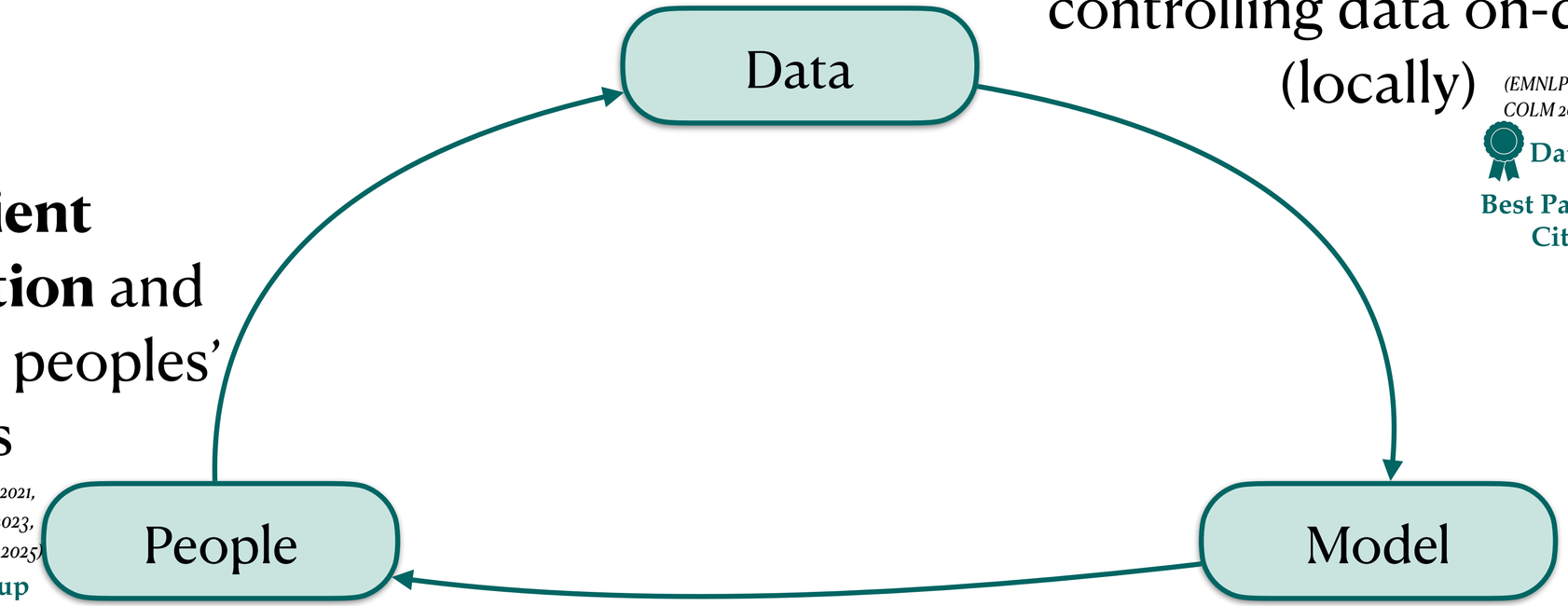
Addressing Violations

(1) **Minimizing** and controlling data on-device (locally)

(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b)

Dataset - 40K Downloads

Best Paper Nominee - Top 20 Most Cited EMNLP 2022 Papers



(3) **Efficient Personalization** and adaptation to peoples' needs

(ASPLOS 2020, WWW 2021, EMNLP 2021, ICIP 2021, NAACL 2021, ACL 2022, NeurIPS 2022, ACL 2023, EMNLP 2023, ICLR 2024, ACL 2024, NAACL 2025)

NCWIT Award, Patent, Startup

(2) **Verification** of contextual and appropriate flow of data by the model

Spotlight *(ICLR2024, EMNLP 2024, COLM 2024)*

Sometimes personal details are needed for accurate predictions!

Example: Medical Query

I'm **34** yo **trans woman** and have been on **oral estradiol** 4 mg/day for three years. My heart suddenly races when I climb stairs and I'm short of breath. What is wrong with me?

[...] Possible causes could be **Pulmonary Embolism (PE)** — a medical emergency, Cardiovascular strain, Respiratory causes or Anemia.

https://chatgpt.com/share/68774952-fa6c-8011-9c79-8dfa2fc8291c

Example: Medical Query, minimized for privacy

I'm **34** yo ~~**trans woman** and have been on **oral estradiol** 4 mg/day for three years.~~ My heart suddenly races when I climb stairs and I'm short of breath. What is wrong with me?

[...] Possible causes could be ~~**Pulmonary Embolism (PE)**~~ — a medical emergency, Cardiovascular strain, Respiratory causes or Anemia.

https://chatgpt.com/share/687747ff-b640-8011-9b0c-59dcc9f0d0da

Example: Medical Query, minimized for privacy

I'm **34** yo ~~**trans woman** and have been on **oral estradiol** 4 mg/day for three years.~~ My heart suddenly races when I climb stairs and I'm short of breath. What is wrong with me?

[...] Possible causes could be ~~**Pulmonary Embolism (PE)**~~ — a medical emergency, Cardiovascular strain, Respiratory causes or Anemia.

The true, serious diagnosis of **Pulmonary Embolism (PE)** is dismissed when sensitive details are removed!

https://chatgpt.com/share/687747ff-b640-8011-9b0c-59dcc9f0d0da

Users prefer low-latency, personalized experiences!

Privacy-Preserving LLM Interaction

with Socratic Chain-of-Thought Reasoning and Homomorphically Encrypted Vector Databases


Yubeen Bae


Minchan Kim


Jaejin Lee


Sangbum Kim


Jaehyung Kim


Yejin Choi

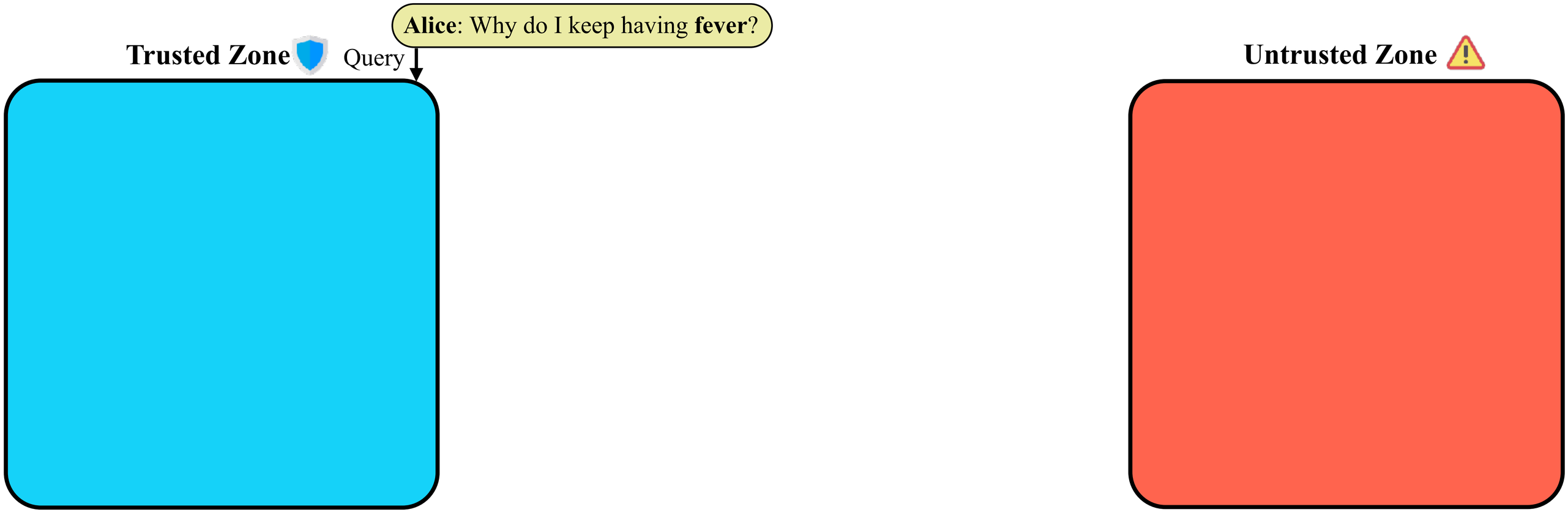

Niloofar Mireshghallah

Socratic Chain-of-Thought Reasoning

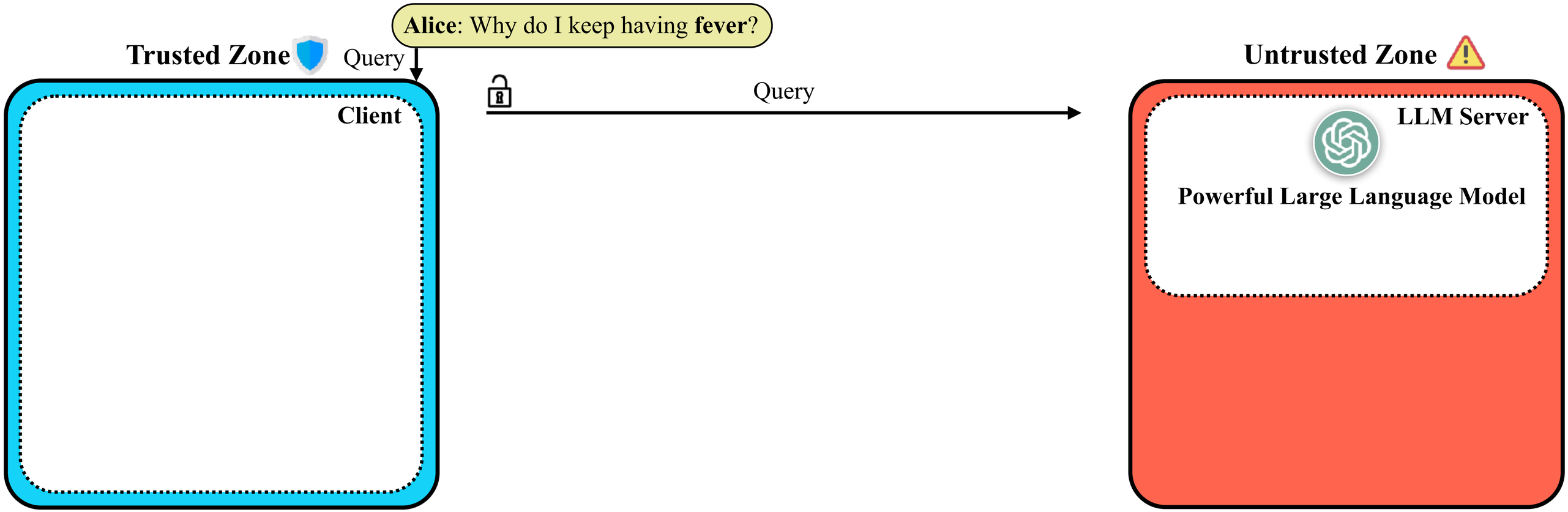
Alice: Why do I keep having **fever**?

Untrusted Zone ⚠️

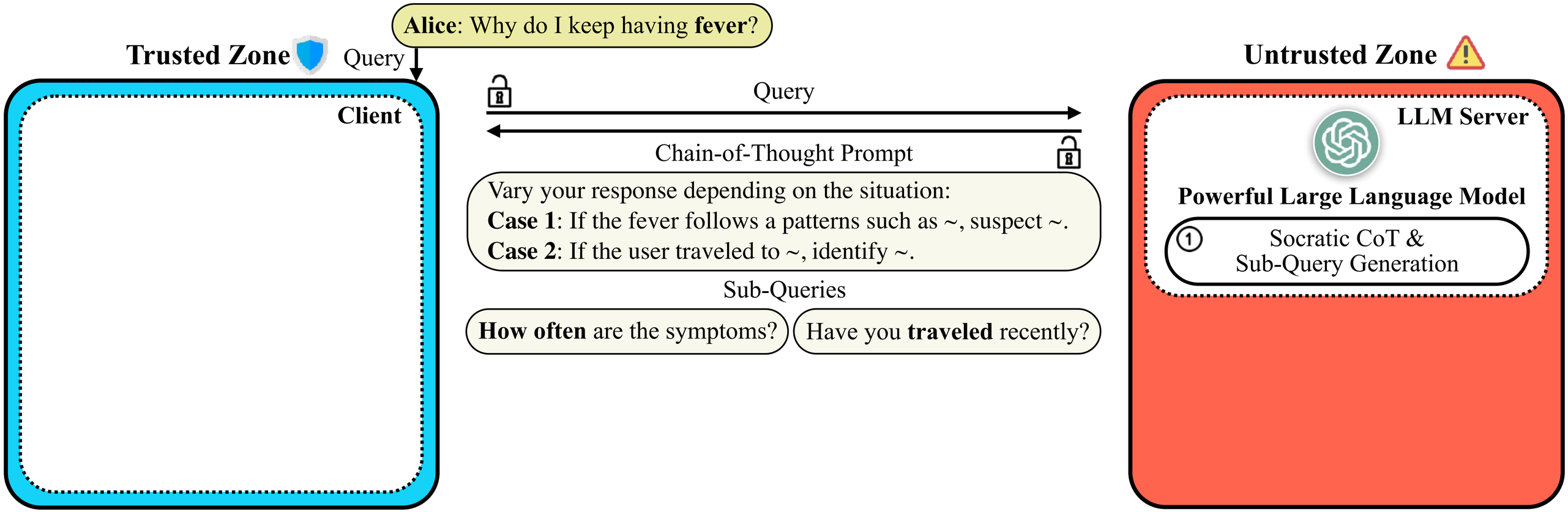
Socratic Chain-of-Thought Reasoning



Socratic Chain-of-Thought Reasoning



Socratic Chain-of-Thought Reasoning



Socratic Chain-of-Thought Reasoning

Alice: Why do I keep having **fever**?

Query

Trusted Zone

Client

Embedding Model

② Sub-Query Embedding Extraction

Query

Chain-of-Thought Prompt

Vary your response depending on the situation:
Case 1: If the fever follows a patterns such as ~, suspect ~.
Case 2: If the user traveled to ~, identify ~.

Sub-Queries

How often are the symptoms?

Have you **traveled** recently?

Untrusted Zone

LLM Server

Powerful Large Language Model

① Socratic CoT & Sub-Query Generation

Socratic Chain-of-Thought Reasoning

Alice: Why do I keep having **fever**?

Query

Trusted Zone

Client

Embedding Model

② Sub-Query Embedding Extraction

Query

Chain-of-Thought Prompt

Vary your response depending on the situation:
Case 1: If the fever follows a patterns such as ~, suspect ~.
Case 2: If the user traveled to ~, identify ~.

Sub-Queries

How often are the symptoms?

Have you **traveled** recently?

Sub-Query Embeddings

Untrusted Zone

LLM Server

Powerful Large Language Model

① Socratic CoT & Sub-Query Generation

DB Server

Alice's Private Database

Storage Offloading

Personal agents need seamless accumulation & real-time retrieval of user data.
Scalable Private Vector Database is needed!

Scalable & Private : Remote Server + Encryption

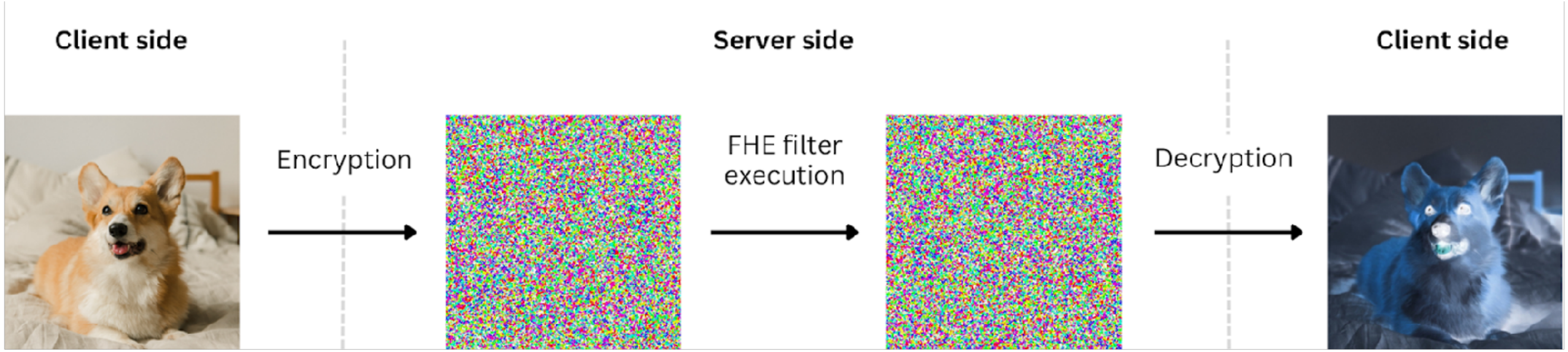
Secure Vector Search over encrypted data : Homomorphic Encryption

→ Optimize cryptographic operations for efficiency

Homomorphically Encrypted Vector Database

Homomorphic Encryption

- Enable operations over encrypted data
  - Operations on the encrypted data are reflected in the underlying data
  - Encrypted data is indistinguishable from noise



https://www.zama.ai/post/encrypted-image-filtering-using-homomorphic-encryption

Homomorphically Encrypted Vector Database

Memory overhead

Latency overhead

Homomorphically Encrypted Vector Database

Memory overhead mitigation

Seeding : Generate a polynomial deterministically from a seed, allowing storage of the seed instead of the full polynomial

MLWE : Reduce the polynomial degree to the dimension of embedding vector

Latency overhead

Homomorphically Encrypted Vector Database

Memory overhead mitigation

Seeding : Generate a polynomial deterministically from a seed, allowing storage of the seed instead of the full polynomial

MLWE : Reduce the polynomial degree to the dimension of embedding vector

Latency overhead mitigation

Cache and Batch the operations that can be precomputed

- Precompute via Key-Query Decoupling
- Additional computation can be reduced by **Butterfly Decomposition**

Socratic Chain-of-Thought Reasoning

Alice: Why do I keep having **fever**?

Query

Trusted Zone

Client

Embedding Model

② Sub-Query Embedding Extraction

Local Language Model

Query

Chain-of-Thought Prompt

Vary your response depending on the situation:
Case 1: If the fever follows a patterns such as ~, suspect ~.
Case 2: If the user traveled to ~, identify ~.

Sub-Queries

How often are the symptoms?

Have you **traveled** recently?

Sub-Query Embeddings

Private Data

Has a fever **every two days**.

Recently traveled to **Africa**.

Untrusted Zone

LLM Server

Powerful Large Language Model

① Socratic CoT & Sub-Query Generation

DB Server

Alice's Private Database

③ Homomorphically Encrypted Vector Similarity Search

Socratic Chain-of-Thought Reasoning

Alice: Why do I keep having **fever**?

Query

Trusted Zone

Client

Embedding Model

② Sub-Query Embedding Extraction

Local Language Model

④ Response Generation via CoT over Private Data

Response

Query

Chain-of-Thought Prompt

Vary your response depending on the situation:
Case 1: If the fever follows a patterns such as ~, suspect ~.
Case 2: If the user traveled to ~, identify ~.

Sub-Queries

How often are the symptoms?

Have you **traveled** recently?

Sub-Query Embeddings

Private Data

Has a fever **every two days**.

Recently traveled to **Africa**.

Untrusted Zone

LLM Server

Powerful Large Language Model

① Socratic CoT & Sub-Query Generation

DB Server

Alice's Private Database

③ Homomorphically Encrypted Vector Similarity Search

Bot: Considering the two-day symptom cycle and recent travel to Africa, **malaria** may be the cause.

Socratic Chain-of-Thought Reasoning

Local-only is enough with relatively simple tasks

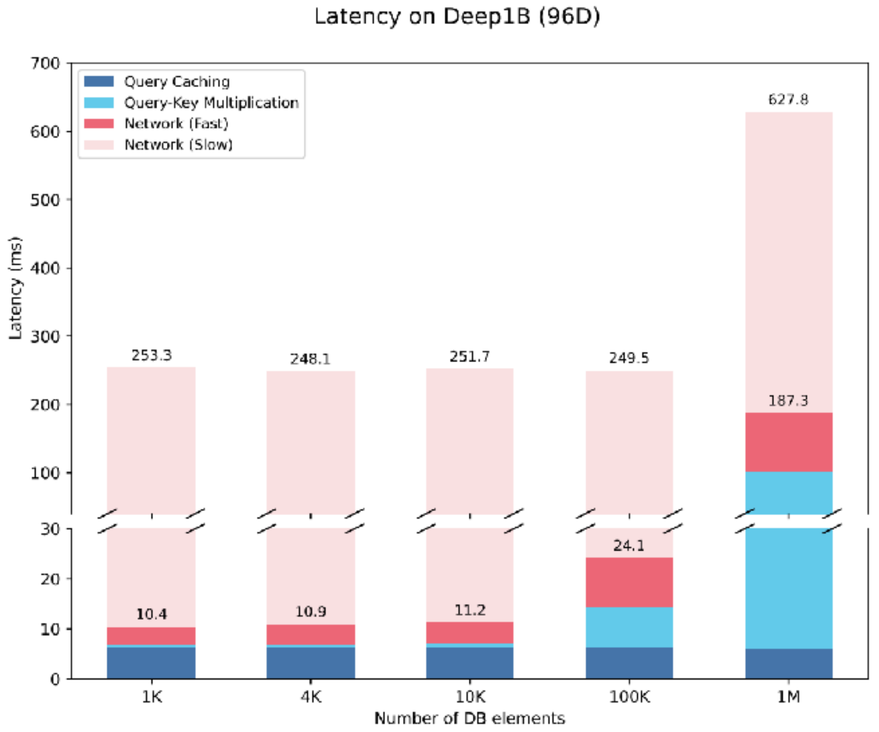
Method	Model	LoCoMo	MediQ
Remote-Only Baseline	R1	80.6	**81.8**
Remote-Only Baseline w/ Socratic CoT	R1 + R1	**92.6**	67.3
Local-Only Baseline	L1	64.6	32.1
Local-Only Baseline w/ Socratic CoT	L1 + L1	82.0	32.5
Hybrid Framwork w/ Socratic CoT (ours)	L1 + R1	87.7	59.7

For casual tasks like LoCoMo, using Socratic CoT on a **single model improves** its performance!

Table 3: The first ablation study for Socratic Chain-of-Thought Reasoning on the **LoCoMo** and **MediQ** datasets. LocoMo is evaluated by F1 score, while MediQ is evaluated by exact match. R1 is GPT-4o, and L1 is Llama-3.2-1B. *Takeaway: Reasoning augmentation through Socratic Chain-of-Thought Reasoning is the primary driver of performance gains.*

Homomorphically Encrypted Vector Databases

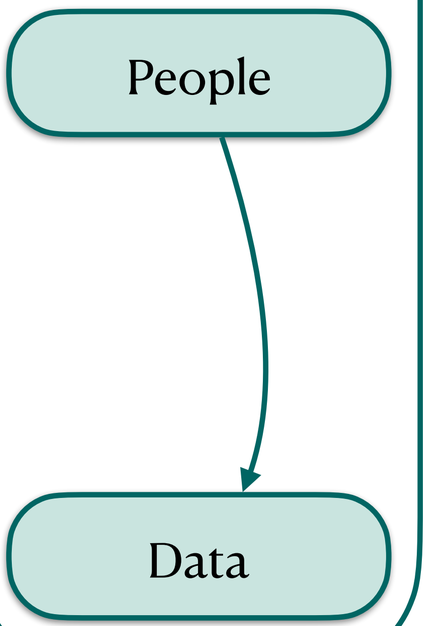
Experiments



Sub-second latency for million scale data!

Figure 2: Multi-thread search latency (using 64 threads) breakdown on the Deep1B [4] dataset as the number of database entries increases. Red and pink bars represent network communication time on fast and slow networks, respectively, while the numbers above each bar indicate the corresponding latency. Blue bars represent query caching time; light-blue bars show query-key multiplication time. *Takeaway: Our encrypted search scales to 1M entries with < 1 second latency, as homomorphic operations incur relatively low overhead compared to network communication.*

Recap



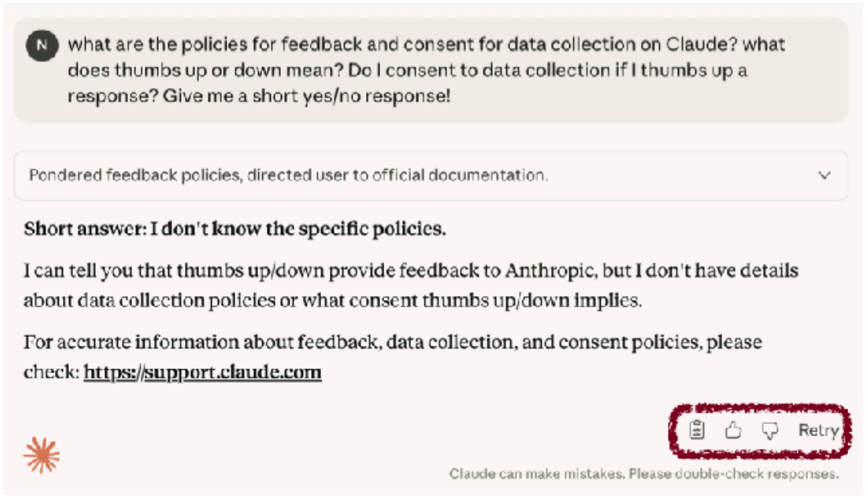
Offloading reasoning + Test time compute: best of both worlds! *(Bae et al. 2025, CRYPTOML)*

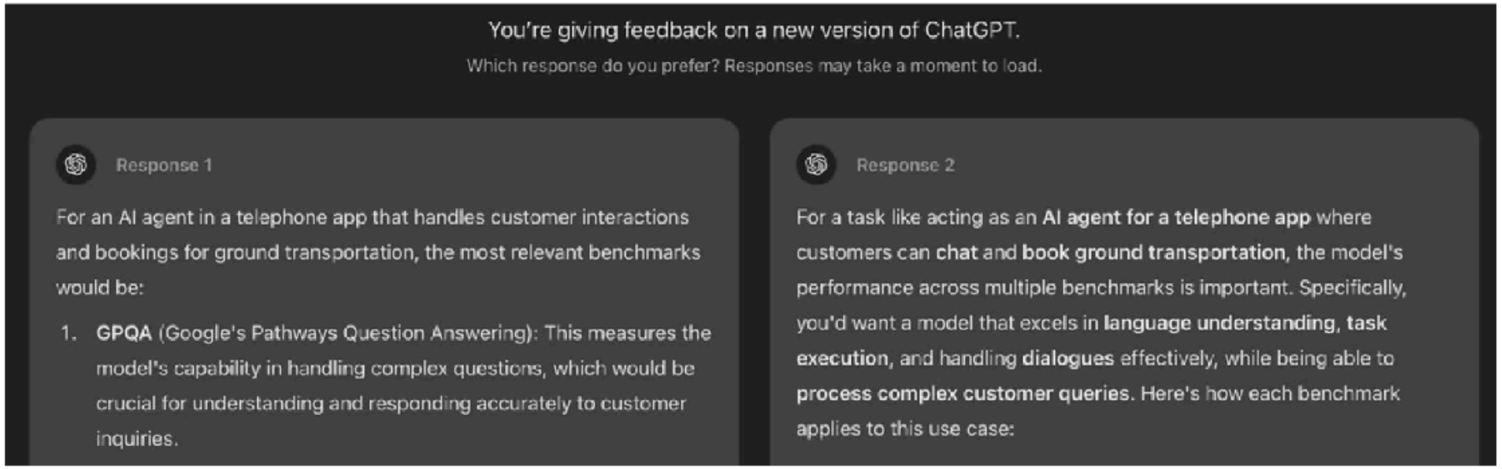
- On-device minimization
- Accuracy restored

Future directions:

- How can we get the local model to perform better using the remote CoTs?
- How do we find the sweet spot of what queries to send and what not to send?

Issues Around Data and Consent





Court Orders

June 5, 2025 Security

How we're responding to The New York Times' data demands in order to protect user privacy

Answers to your questions

Why are The New York Times and other plaintiffs asking for this?

- The New York Times is suing OpenAI. As part of their baseless lawsuit, they've recently asked the court to force us to retain all user content indefinitely going forward, based on speculation that they might find something that supports their case.
- We strongly believe this is an overreach. It risks your privacy without actually helping resolve the lawsuit. That's why we're fighting it.

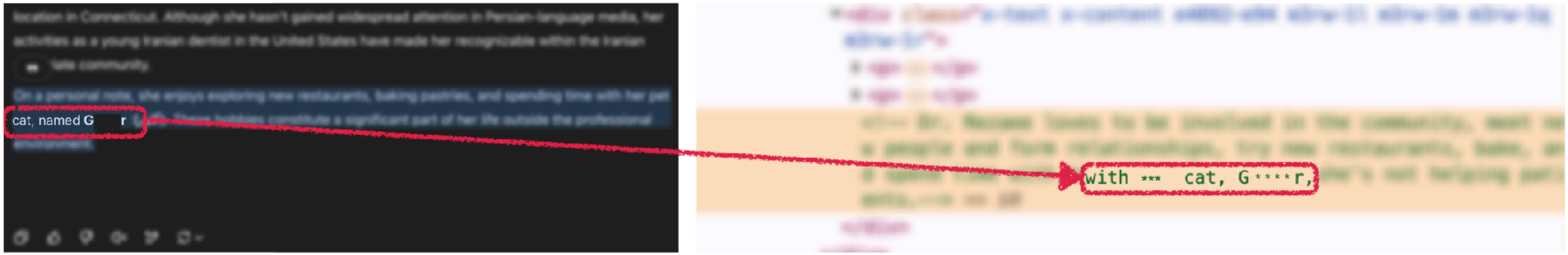
Is my data impacted?

- Yes, if you have a ChatGPT Free, Plus, Pro, and Team subscription or if you use the OpenAI API (without a Zero Data Retention agreement).
- This does not impact ChatGPT Enterprise or ChatGPT Edu customers.
- This does not impact API customers who are using Zero Data Retention endpoints under our ZDR amendment.

https://openai.com/index/response-to-nyt-data-demands/

LLMs as Search Engines and Aggregators

Inferring attributes



These are secondary questions asked for password recovery!

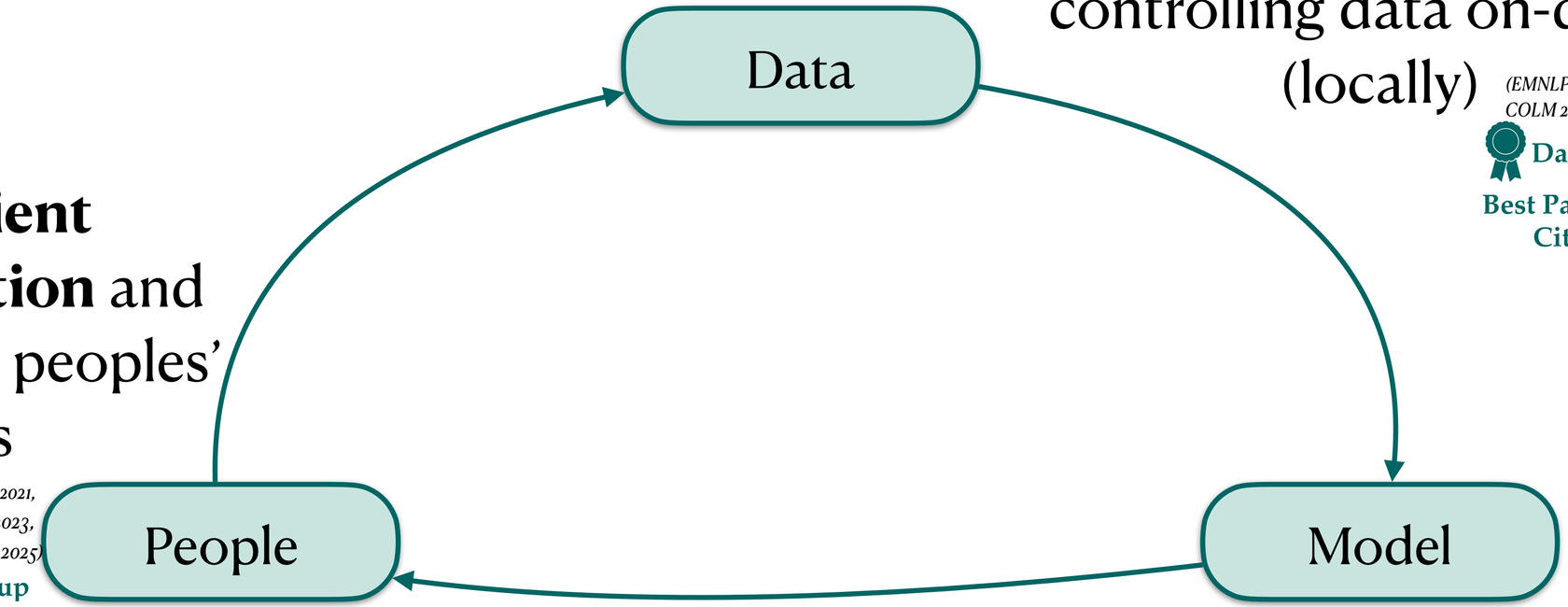
Addressing Violations

(1) **Minimizing** and controlling data on-device (locally)

(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b)

Dataset - 40K Downloads

Best Paper Nominee - Top 20 Most Cited EMNLP 2022 Papers



(3) **Efficient Personalization** and adaptation to peoples' needs

(ASPLOS 2020, WWW 2021, EMNLP 2021, ICIP 2021, NAACL 2021, ACL 2022, NeurIPS 2022, ACL 2023, EMNLP 2023, ICLR 2024, ACL 2024, NAACL 2025)

NCWIT Award, Patent, Startup

(2) **Verification** of contextual and appropriate flow of data by the model

Spotlight *(ICLR2024, EMNLP 2024, COLM 2024)*

Thank You!

niloofar@cmu.edu