

Operationalizing Data Minimization for Privacy-Preserving LLM Prompting



Jijie Zhou



Niloofar Miresghallah



Tianshi Li



Recording available on [SlidesLive](#)

Oversharing in LLM Prompting

I'll travel from **Boston** to Rio on **April 26**.
Please plan my trip itinerary there in **Brazil**.

I'll travel from Boston to Rio on April 26.
Please plan my trip itinerary there in Brazil.

Rio is an amazing destination 🌳 — beaches, mountains, music, and iconic landmarks. Since you're traveling from **Boston** to **Rio de Janeiro** on **April 26**, I'll suggest a **6-day itinerary**, which works well for most first visits. You can easily shorten or extend it.

🌳 Rio de Janeiro Travel Itinerary (6 Days)

Day 1 — Arrival & Beach Sunset



Morning / Afternoon

- Arrive at **Rio de Janeiro–Galeão International Airport**
- Check into your hotel in **Copacabana** or **Ipanema**

Afternoon

- Walk along **Copacabana Beach**
- Grab coconut water or açai from beach vendors

Evening

- Watch sunset at **Arpoador**
- Dinner at **Garota de Ipanema** (famous for Brazilian dishes)

Day 2 — Christ the Redeemer & Santa Teresa



I'll travel from **a city in the US** to Rio in **late April**.
Please plan my trip itinerary.

I'll travel from a city in US to Rio in late April.
Please plan my trip itinerary.

Rio is a fantastic destination—late April is actually a **great time** (warm weather, fewer crowds than peak summer). Below is a **balanced 5-day itinerary** that mixes iconic sights, beaches, culture, and food. If you want, I can later customize it based on your travel style (luxury, nightlife, nature, photography, etc.). 🌳

Rio de Janeiro Trip Itinerary (5 Days)

Day 1 — Arrival & Sunset at Sugarloaf



Morning / Afternoon

- Arrive at **Rio de Janeiro**
- Check into hotel in **Copacabana** or **Ipanema**
- Walk along **Copacabana Beach**

Late Afternoon

- Take the cable car to **Sugarloaf Mountain**
- One of the **best sunset views in the world**

Dinner

- Try Brazilian barbecue at **Fogo de Chão**

Day 2 — Christ the Redeemer & Santa Teresa



Research Gap

Heuristic Prompt Sanitization

- Detect PII → apply rules
(*Dou et al., 2024; Zeng et al., 2025*)
- Heuristic sensitive info detection
(*Chowdhury et al., 2025*)
- LLM-as-a-judge for relevance
(*Ma et al., 2025; Ngong et al., 2025*)

Privacy-Utility Tradeoff Frameworks

- seek a balance between privacy and utility (*Li et al., 2025*)
- aim to maximize utility under a privacy constraint

Training-time Defenses

- Differential privacy training
- Machine unlearning
(*Abadi et al., 2016; Barbulescu & Triantafillou, 2024*)

⚠ Assumes LLMs can predict what information is necessary

⚠ Assumes users accept utility degradation

⚠ Requires model access & trusted providers

Our Goal: Find the most privacy-preserving prompt that still preserves task utility.

Data Minimization

$$\max_{a \in A^n} \text{Priv}(\tau(x; a))$$

Input prompt
Input prompt -> Sanitized input prompt
Action space:
Redact > Abstract > Retain

subject to

$$\text{Util}(\mathcal{R}(\mathcal{F}(\tau(x; a)); a)) \geq \gamma$$

Response quality threshold
Output of the sanitized input prompt
Restored output of the sanitized input prompt

Pipeline

Original Prompt



I want you to act as my travel agent and help me plan an itinerary for a trip to **San Francisco** and **Yosemite National Park** in **California**. We have already booked flights from **Houston** to **San Francisco** for our onward journey on **January 25th** and return on **January 28th**. There are four of us, and we're planning to spend two days in **San Francisco** and one day in **Yosemite National Park**. I'd like your help with planning the itinerary.

Pipeline

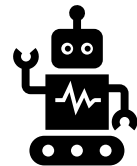
Stage 1: Probe single-PII; mark utility-breaking entities as frozen.



I want you to act as my travel agent and help me plan an itinerary for a trip to [GEOLOCATION1] and [GEOLOCATION2] in California. We have already booked flights from Houston to San Francisco for our onward journey on January 25th and return on January 28th. There are four of us, and we're planning to spend two days in [GEOLOCATION1] and one day in [GEOLOCATION2]. I'd like your help with planning the itinerary.

San Francisco: Retain
Yosemite National Park: Redact
California: Retain
Houston: Retain
January 25th: Retain
January 28th: Retain

Sorry, I need more info about.... :(



Utility check: FAIL

Pipeline

Stage2: Priority-queue tree search on PII combinations for maximal privacy.

0

San Francisco: **Abstract**
Yosemite National Park: **Abstract**
California: **Redact**
Houston: **Redact**
January 25th: **Redact**
January 28th: **Redact**

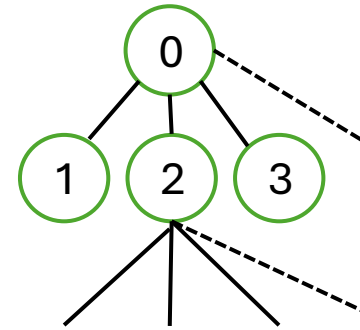
1

San Francisco: **Retain**
Yosemite National Park: **Abstract**
California: **Redact**
Houston: **Redact**
January 25th: **Redact**
January 28th: **Redact**

Privacy Ranking Priority Queue



San Francisco: **Abstract**
Yosemite National Park: **Retain**
California: **Redact**
Houston: **Redact**
January 25th: **Redact**
January 28th: **Redact**



3

San Francisco: **Abstract**
Yosemite National Park: **Abstract**
California: **Abstract**
Houston: **Redact**
January 25th: **Redact**
January 28th: **Redact**

Utility check: **FAIL**

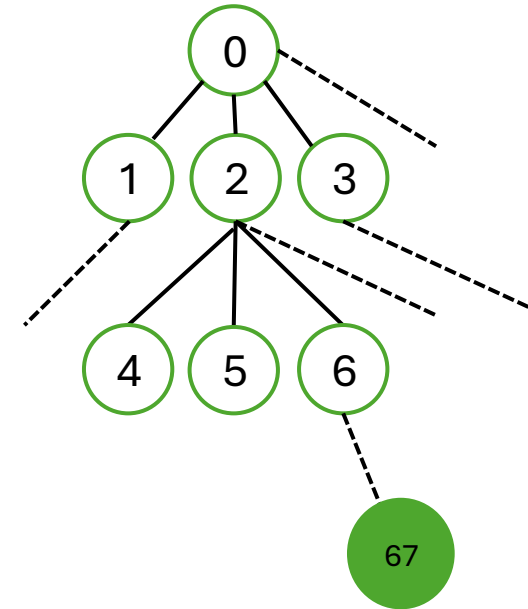
Pipeline

Stage2: Priority-queue tree search on PII combinations for maximal privacy.

67

San Francisco: **Abstract**
Yosemite National Park: **Retain**
California: **Redact**
Houston: **Abstract**
January 25th: **Redact**
January 28th: **Redact**

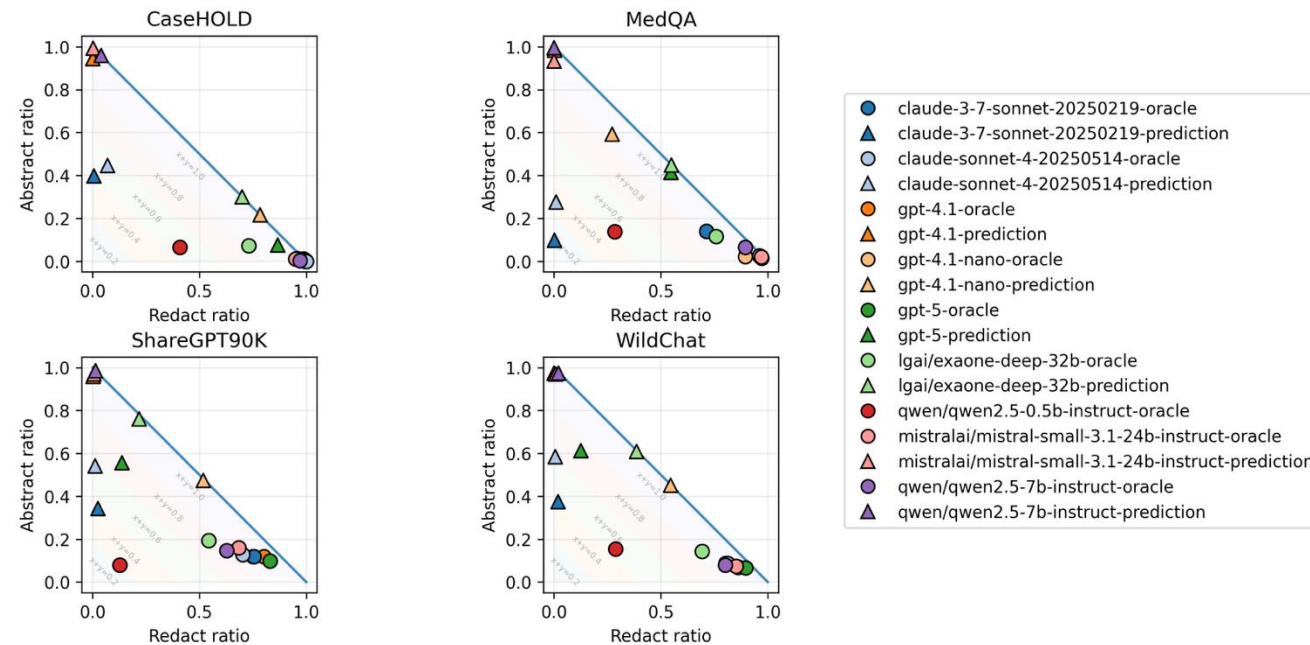
Privacy Ranking Priority Queue



Utility check: **PASS**

Results

- [Response Generation] More capable models require less user information to solve the same task.
- [Data Minimization Prediction] Models overshare relative to the optimal level of minimization.



Contributions

- Formalize data minimization for LLM prompting
- Search algorithm to identify minimal privacy-utility tradeoff
- Larger frontier LLMs can tolerate stronger data minimization while maintaining task quality than smaller opensource models
- Models may lack awareness of what information they actually need to solve a task.

