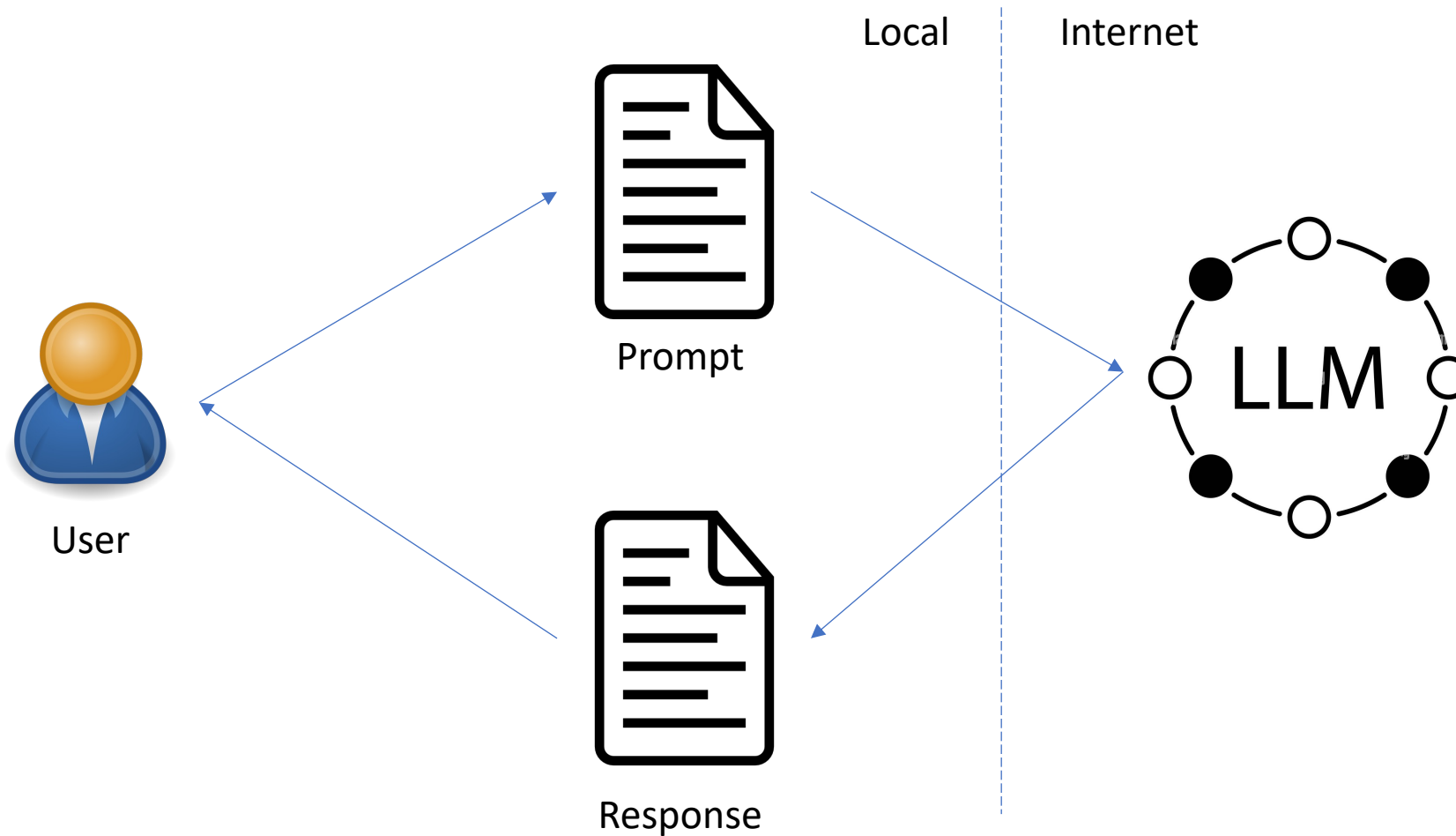


# Privacy Protection in Large Language Model Prompts

Youxiang Zhu

# When we use online/API-based LLMs



# You may leak your privacy from your prompts!

My name is **Tom**, please write an email for me to ask for a leave for tomorrow's **Cybersecurity in the Internet of Things** class.....



User



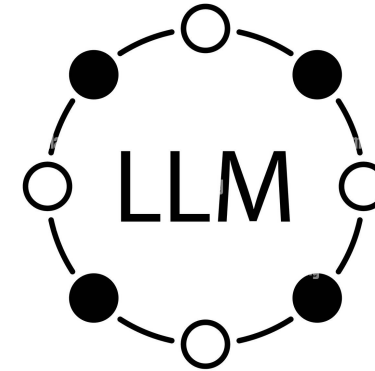
Prompt



Response

Local

Internet



# Privacy leakage in the prompt is severe!

- From the 570k real-world user-ChatGPT interactions in the WildChat dataset, we identified around **59.67 %** of the user prompt may contain some sort of privacy!

The screenshot shows the Hugging Face dataset page for 'WildChat' by allenai. The dataset has 1,387 likes and 121 follows. It is categorized under 'Text Generation', 'Question Answering', and 'Text2Text Generation'. The dataset is in 'parquet' format, with a size of '100K - 1M' and is licensed under 'odc-by'.

The 'Dataset Viewer' section shows the dataset is split into 'train' with 529k rows. It includes search and filter options, and a table of data rows. The table columns are 'conversation\_id', 'model', 'timestamp', 'conversation', and 'turn'. The 'conversation' column contains JSON objects with 'content', 'language', 'redacted', and 'role' fields.

conversation_id	model	timestamp	conversation	turn
26c5dc109920789f9199ff9b37acb8c1	gpt-4	"2023-04-10T00:01:08"	[ { "content": "Write a very long, elaborate, descriptive and detailed..."	1
e87a1aeb9aafa35c00da39ddeb1139a0	gpt-4	"2023-04-10T00:01:10"	[ { "content": "what are you?", "language": "English", "redacted": false, "role": "...	1
c3415a9e401ff379f29fe3ce02e500dc	gpt-4	"2023-04-10T00:02:37"	[ { "content": "Write an engaging and a constructive article for my Morocco travel..."	1
ec6578cd9d69130a769ea307b6e7a874	gpt-4	"2023-04-10T00:03:07"	[ { "content": "CONSTRAINTS:\n\n1. ~4000 word limit for short term memory. Your..."	1
17827951de7dc8b29e8e2baa1a73e875	gpt-3.5-turbo	"2023-04-10T00:03:09"	[ { "content": "اكتب لي بحث عن اعميه نظام الاتحاد النقل الجوي الدولي بالنسبة لاشخاص..."	3
c9fdeddb0e222c9251e5fab2b0784240c	gpt-4	"2023-04-10T00:03:58"	[ { "content": "Write an engaging and a constructive article for my Morocco travel..."	1

# Two ways to protect privacy in the prompts

- Model centric
  - Use local models to hide privacy information before API use, and add privacy information back in the response
  - Pros: Fully automatic
  - Cons: hard to ensure the quality for specific prompts
- User centric
  - Use local models to inform the user about the privacy in the prompts and the utility impact to hide them
  - Pros: user can choose case by case based on their need
  - Cons: need human intervention

# Two ways to protect privacy in the prompts

Model centric:  
concurrent work from  
Columbia University

**PAPILLON : PrivAcy Preservation from Internet-based and Local Language Model Ensembles**

**Li Siyan<sup>1</sup>, Vethavikashini Chithrara Raghuram<sup>1</sup>, Omar Khattab<sup>2,3</sup>,  
Julia Hirschberg<sup>1</sup>, Zhou Yu<sup>1</sup>**

<sup>1</sup>Columbia University, <sup>2</sup>Stanford University, <sup>3</sup>Databricks

Correspondence: [siyan.li@columbia.edu](mailto:siyan.li@columbia.edu)

User centric:  
Our work

**Exploiting Privacy Preserving Prompt Techniques  
for Online Large Language Model Usage**

Youxiang Zhu, Ning Gao, Xiaohui Liang, and Honggang Zhang

Department of Computer Science

University of Massachusetts Boston, MA, USA

Email: {Youxiang.Zhu001, Ning.Gao001, Xiaohui.Liang, Honggang.Zhang}@umb.edu

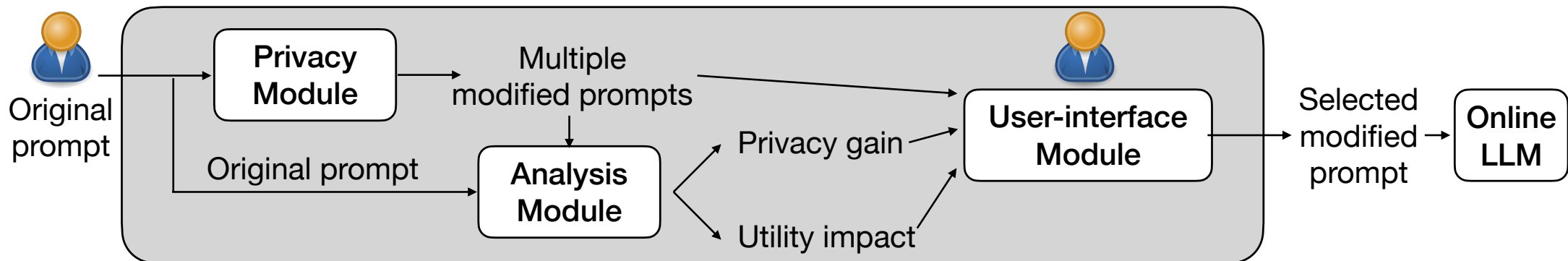
# Goals of our user centric approach

- Identify privacy information in the prompt
- Inference utility impact if some privacy information is hided
- Inform users about the privacy information and utility impact

Privacy module

Analysis module

User-interface module



# How to identify privacy information

- Named entity recognition (NER)

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's **PaperAdvertisementSupported** **ORG** by F.B.I. Agent **Peter Strzok** **PERSON**,  
**Who Criticized Trump** **PERSON** in Texts, Is FiredImagePeter Strzok, a top **F.B.I.** **GPE** counterintelligence agent who was taken off the special counsel  
investigation after his disparaging texts about President **Trump** **PERSON** were uncovered, was fired. **CreditT.J. Kirkpatrick** **PERSON** for **The New York**  
**TimesBy Adam Goldman** **ORG** and **Michael S. SchmidtAug** **PERSON**. **13** **CARDINAL**, **2018WASHINGTON** **CARDINAL** — **Peter Strzok**  
**PERSON**, the **F.B.I.** **GPE** senior counterintelligence agent who disparaged President **Trump** **PERSON** in inflammatory text messages and helped  
oversee the **Hillary Clinton** **PERSON** email and **Russia** **GPE** investigations, has been fired for violating bureau policies, Mr. **Strzok** **PERSON**'s lawyer  
said **Monday** **DATE**. Mr. Trump and his allies seized on the texts — exchanged during the **2016** **DATE** campaign with a former **F.B.I.** **GPE** lawyer,  
**Lisa Page** — in **PERSON** assailing the **Russia** **GPE** investigation as an illegitimate “witch hunt.” Mr. **Strzok** **PERSON**, who rose over **20 years**  
**DATE** at the **F.B.I.** **GPE** to become one of its most experienced counterintelligence agents, was a key figure in **the early months** **DATE** of the  
inquiry. Along with writing the texts, Mr. **Strzok** **PERSON** was accused of sending a highly sensitive search warrant to his personal email account. The  
**F.B.I.** **GPE** had been under immense political pressure by Mr. **Trump** **PERSON** to dismiss Mr. **Strzok** **PERSON**, who was removed **last summer**  
**DATE** from the staff of the special counsel, **Robert S. Mueller III** **PERSON**. The president has repeatedly denounced Mr. **Strzok** **PERSON** in posts on



# How to identify privacy information

- Named entity recognition (NER)
- Training set

**Datasets:** 4 ai4privacy/pii-masking-200k 83 Follow 4 Ai4Privacy 21

Tasks: Text Classification Token Classification Table Question Answering +13 Modalities: Text Formats: json Languages:

Tags: legal business psychology privacy DOI: doi:10.57967/hf/1532 Libraries: Datasets Disk Croissant +1

Dataset card Viewer Files and versions Community 11

---

**Dataset Viewer** Auto-converted to Parquet API Embed Full Screen Viewer

Split (1)  
train · 209k rows

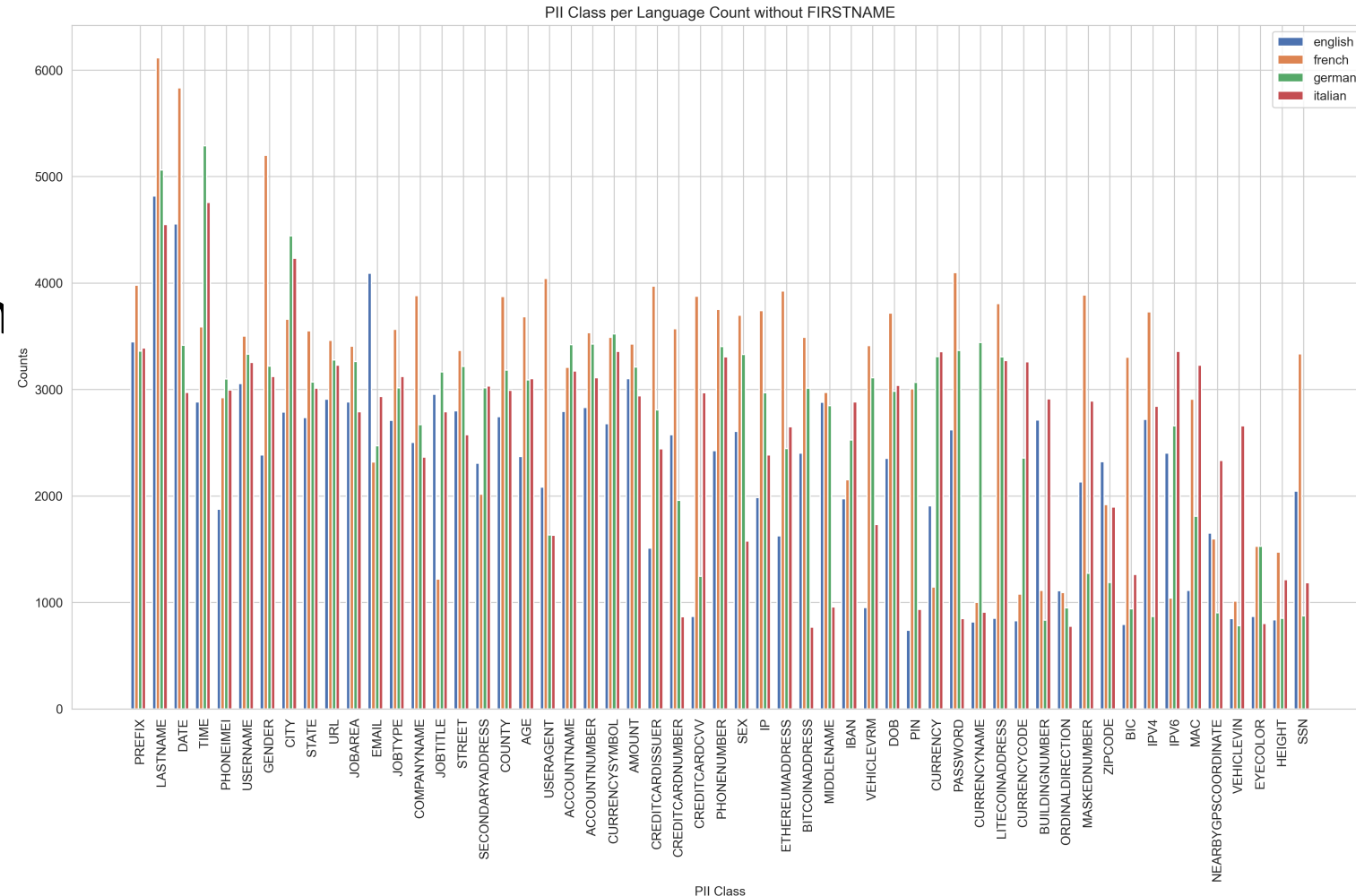
Search this dataset SQL Console

source_text string · lengths	target_text string · lengths	privacy_mask list · lengths	span_labels string · lengths
 31 2.79k	 41 2.71k	 2 42	 49 1.84k
A student's assessment was found on device bearing IMEI: 06-184755-866851-3. The...	A student's assessment was found on device...	[ { "value": "06-184755-866851-3", "start": 57, "end": 75, "label": "PHONEIMEI" }, { ...	[[0, 57, "0"], [57, "PHONEIMEI"], [75, 1
Dear Omer, as per our records, your license 78B5R2MVFAHJ48500 is still registered in...	Dear [FIRSTNAME], as per our records, your...	[ { "value": "Omer", "start": 5, "end": 9, "label": "FIRSTNAME" }, { "value": "...	[[0, 5, "0"], [5, 9, "FIRSTNAME"], [9, 44
Kattie could you please share your recommendations about vegetarian diet for 7...	[FIRSTNAME] could you please share your...	[ { "value": "Kattie", "start": 0, "end": 6, "label": "FIRSTNAME" }, { "value": "72", ...	[[0, 6, "FIRSTNAME"], [6, 75, "0"], [75, 7
Emergency supplies in 16356 need a refill. Use 5890724654311332 to pay for them.	Emergency supplies in [BUILDINGNUMBER] need...	[ { "value": "16356", "start": 22, "end": 27, "label": "BUILDINGNUMBER" }, { "value": "...	[[0, 22, "0"], [22, "BUILDINGNUMBER"], [
The 88 old child at 5862, has showcased an unusual ability to remember and recite...	The [AGE] old child at [BUILDINGNUMBER], has...	[ { "value": "88", "start": 4, "end": 6, "label": "AGE" }, { "value": "5862", "start": ...	[[0, 4, "0"], [4, 6, "AGE"], [6, 20, "0"]
Your recent hospital data recorded on 29/12/1957 regarding chronic disease...	Your recent hospital data recorded on [DOB]...	[ { "value": "29/12/1957", "start": 38, "end": 48, "label": "DOB" }, { "value": "...	[[0, 38, "0"], [38, "DOB"], [48, 115, "C

< Previous 1 2 3 ... 2,093 Next >

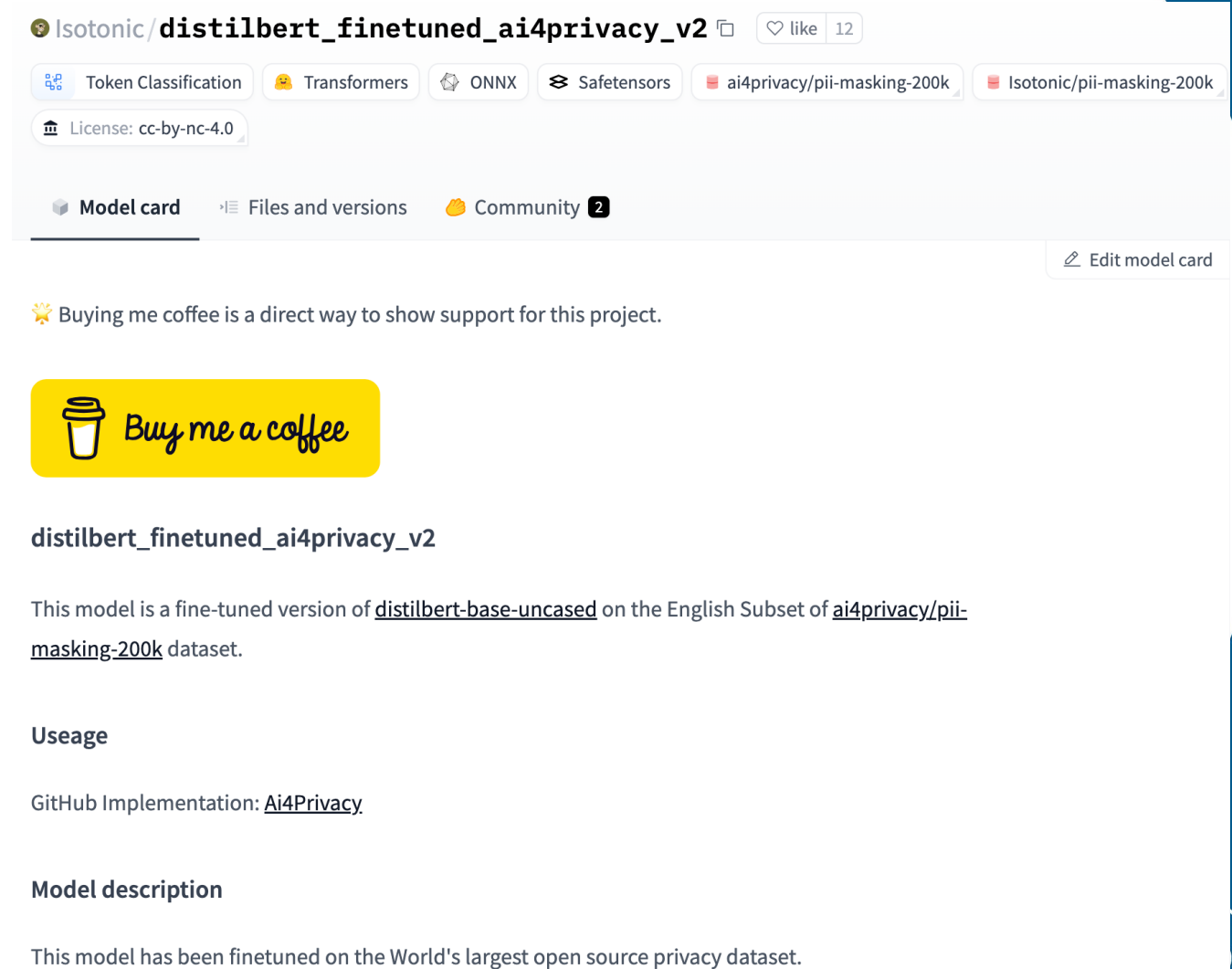
# How to identify privacy information

- Named entity recognition (NER)
- Training set
- 54 pre-defined PII (Person Identifiable Information) classes



# How to identify privacy information

- Named entity recognition (NER)
- Models



The screenshot shows the Hugging Face model card for **Isotonic/distilbert\_finetuned\_ai4privacy\_v2**. The card includes a header with the model name, a 'like' button with 12 likes, and a row of tags: Token Classification, Transformers, ONNX, Safetensors, ai4privacy/pii-masking-200k, and Isotonic/pii-masking-200k. Below the tags is the license: cc-by-nc-4.0. The main content area has tabs for 'Model card' (selected), 'Files and versions', and 'Community'. A yellow 'Buy me a coffee' button is prominently displayed. The description states: 'This model is a fine-tuned version of [distilbert-base-uncased](#) on the English Subset of [ai4privacy/pii-masking-200k](#) dataset.' The 'Usage' section mentions a GitHub implementation at [Ai4Privacy](#). The 'Model description' section states: 'This model has been finetuned on the World's largest open source privacy dataset.'

Isotonic/**distilbert\_finetuned\_ai4privacy\_v2** like 12


Token Classification Transformers ONNX Safetensors ai4privacy/pii-masking-200k Isotonic/pii-masking-200k

License: cc-by-nc-4.0

Model card Files and versions Community 2

Edit model card

☕ Buying me coffee is a direct way to show support for this project.

 Buy me a coffee

**distilbert\_finetuned\_ai4privacy\_v2**

This model is a fine-tuned version of [distilbert-base-uncased](#) on the English Subset of [ai4privacy/pii-masking-200k](#) dataset.

**Usage**

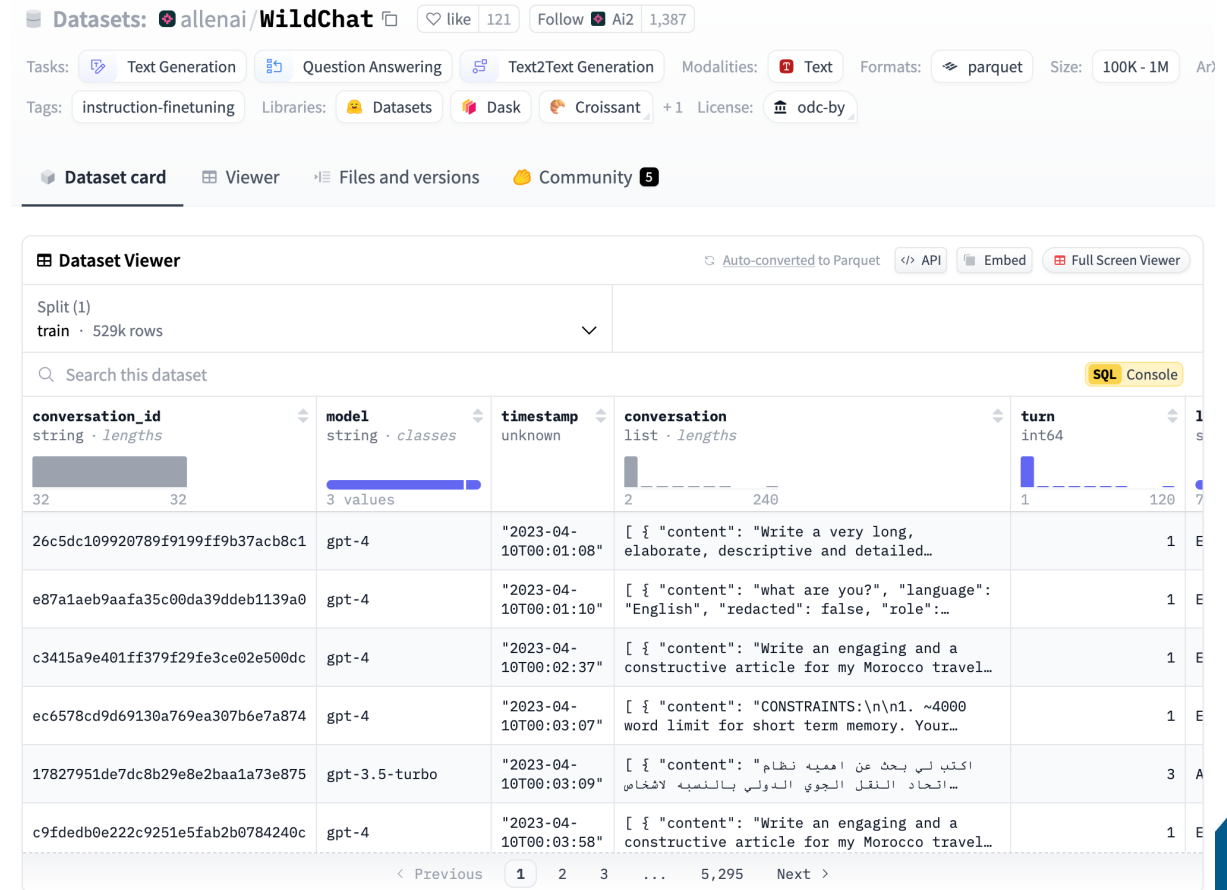
GitHub Implementation: [Ai4Privacy](#)

**Model description**

This model has been finetuned on the World's largest open source privacy dataset.

# How to identify privacy information

- Named entity recognition (NER)
- Apply NER model to the Wildchat dataset (570K prompts)
- Data pre-processing: For each of 54 PII class, sample a maximum of 200 prompts, results in 7623 original prompts in total.



**Datasets:** allenai/WildChat like 121 Follow Ai2 1,387

Tasks: Text Generation Question Answering Text2Text Generation Modalities: Text Formats: parquet Size: 100K - 1M Ar

Tags: instruction-finetuning Libraries: Datasets Dask Croissant +1 License: odc-by

**Dataset card** **Viewer** **Files and versions** **Community 5**

**Dataset Viewer** Auto-converted to Parquet </> API Embed Full Screen Viewer

Split (1)  
train · 529k rows

Search this dataset SQL Console

conversation_id string · lengths	model string · classes	timestamp unknown	conversation list · lengths	turn int64
26c5dc109920789f9199ff9b37acb8c1	gpt-4	"2023-04-10T00:01:08"	[ { "content": "Write a very long, elaborate, descriptive and detailed..."	1
e87a1aeb9aafa35c00da39ddeb1139a0	gpt-4	"2023-04-10T00:01:10"	[ { "content": "what are you?", "language": "English", "redacted": false, "role":...	1
c3415a9e401ff379f29fe3ce02e500dc	gpt-4	"2023-04-10T00:02:37"	[ { "content": "Write an engaging and a constructive article for my Morocco travel..."	1
ec6578cd9d69130a769ea307b6e7a874	gpt-4	"2023-04-10T00:03:07"	[ { "content": "CONSTRAINTS:\n\n1. ~4000 word limit for short term memory. Your..."	1
17827951de7dc8b29e8e2baa1a73e875	gpt-3.5-turbo	"2023-04-10T00:03:09"	[ { "content": "اكتب لي بحث عن اهميه نظام اتحاد النقل الجوي الدولي بالنسبه لاشخاص..."	3
c9fdedb0e222c9251e5fab2b0784240c	gpt-4	"2023-04-10T00:03:58"	[ { "content": "Write an engaging and a constructive article for my Morocco travel..."	1

< Previous 1 2 3 ... 5,295 Next >

# How to hide the privacy information

Original prompts

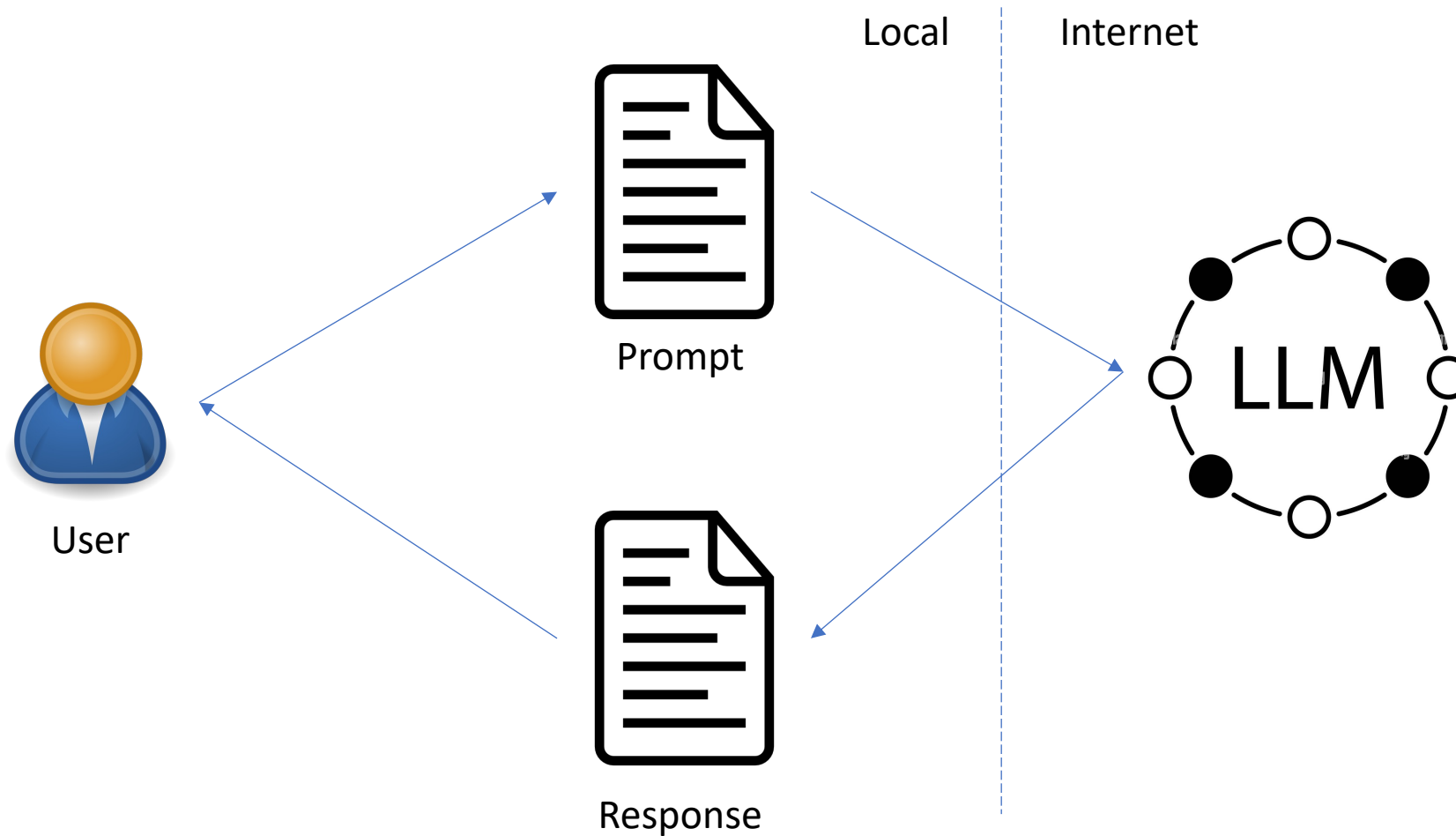


- **Four privacy techniques** (e.g., for prompt “why do sales engineering managers like their job?”)
- **Remove** the keyword: “why do like their job?” ← **Modified prompts**
- **Mask** the keyword with category: “why do [jobtitle] like their job?”
- **Replace** the keyword with another keyword in the same category: “why do Regional Brand Analysts like their job?”
- **Rewrite** the prompt to hide keyword with a local LLM: “Why do engineering managers like their job:”

Modified prompts

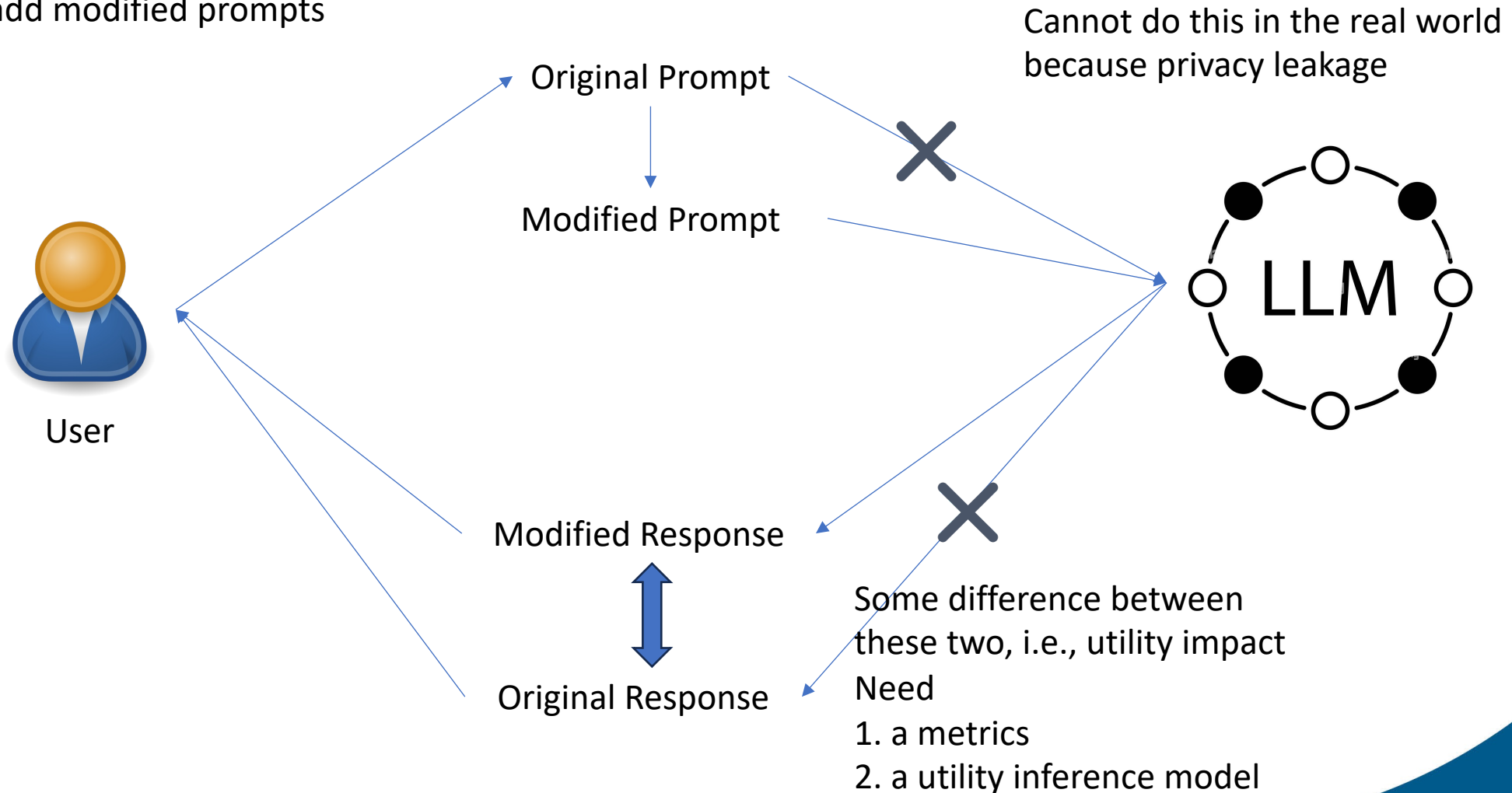


# How to measure and infer the utility impact



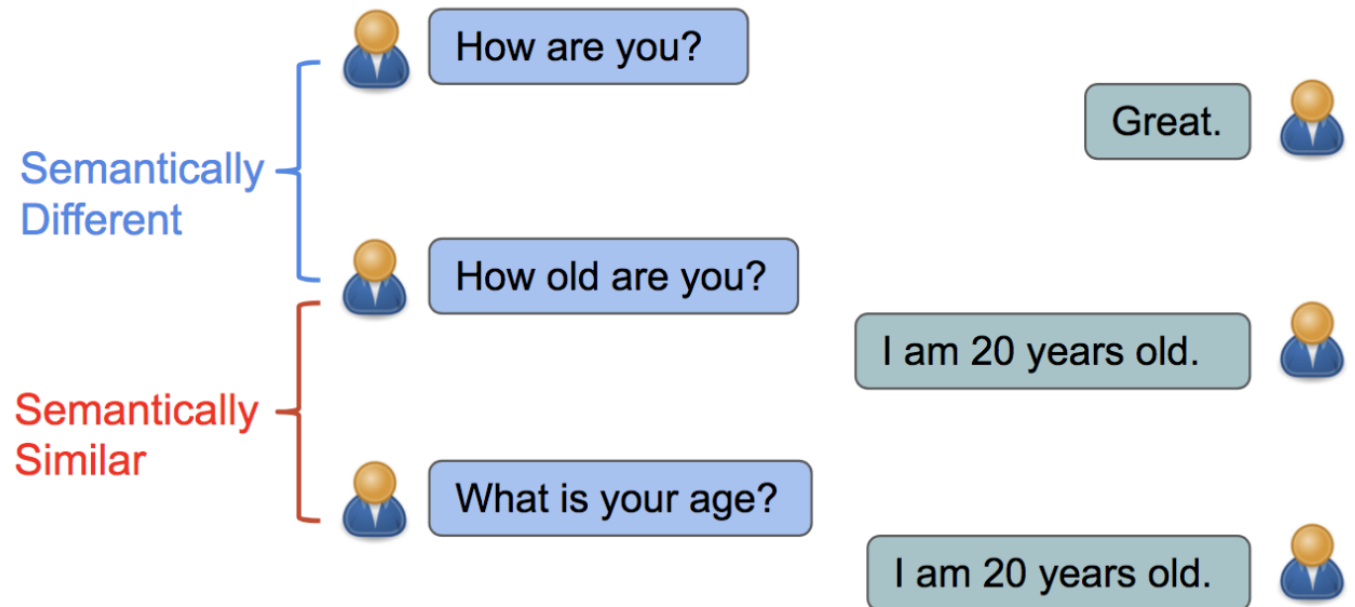
# How to measure and infer the utility impact

If we add modified prompts



# How to measure and infer the utility impact

- Semantic similarity



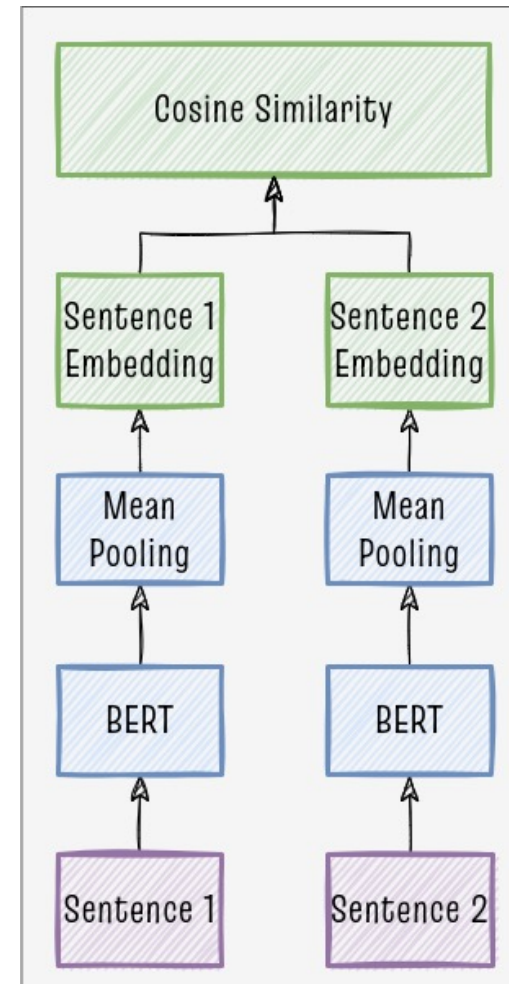


# How to measure and infer the utility impact

- Semantic similarity
- Semantic similarity models

We don't have original response in the real world

Original response



Similarity(Original response, modified response)

Modified response

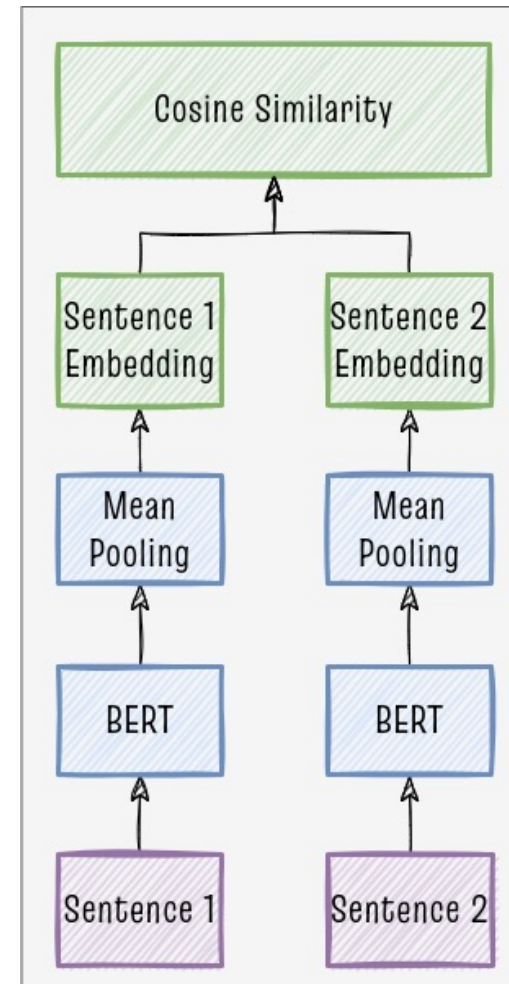
# How to measure and infer the utility impact

- Semantic similarity
- Utility inference model

Need to use prompts to predict similarity in the response!

Original prompt

Original ~~response~~



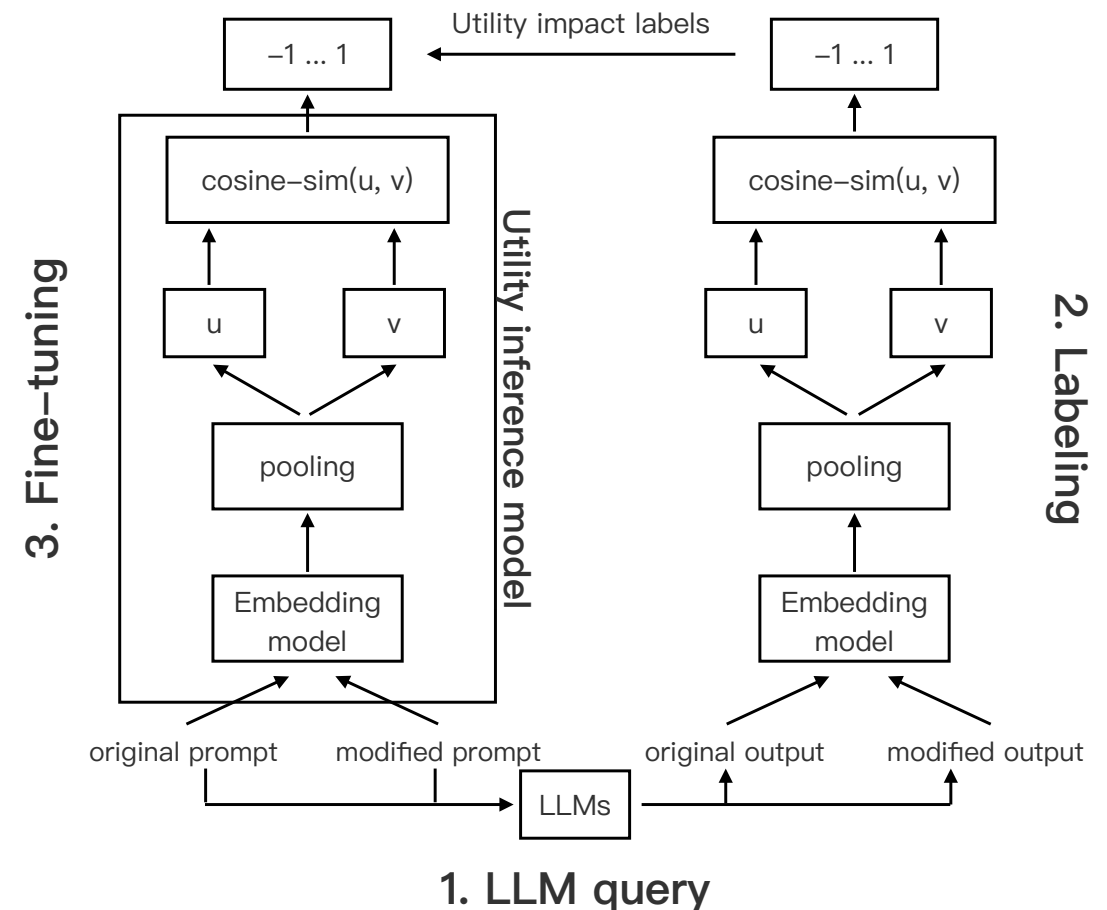
Similarity(Original  
response, modified  
response)

Modified prompt

Modified ~~response~~

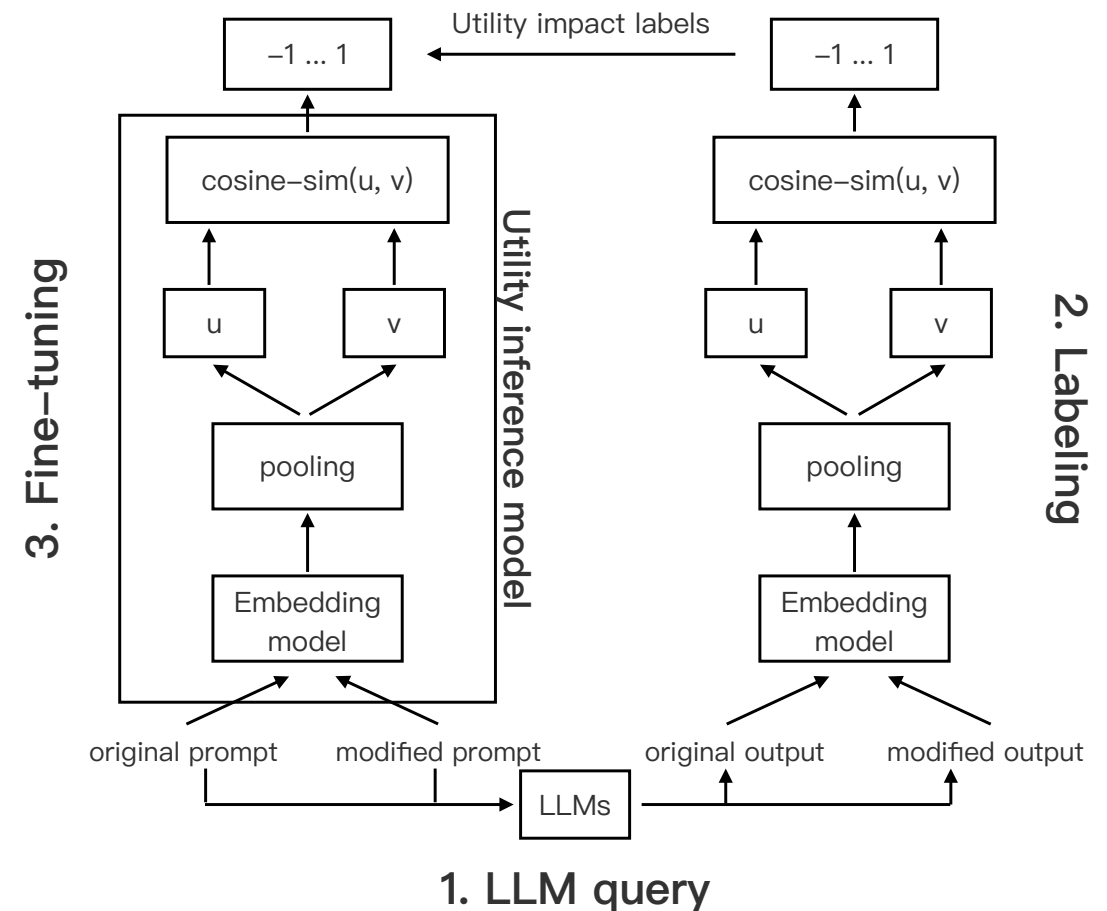
# How to measure and infer the utility impact

- **LLM query:** Apply four privacy techniques to obtain the modify prompts. Query the Llama 2 7B model to obtain the LLM outputs for both original prompts and modified prompts.
- **Labeling:** Define utility impact as semantic similarity between original output and modified output. Obtain the utility impact label using an embedding model.
- **Fine-tuning:** Fine tune the embedding model with the utility impact label, obtaining utility inference model.



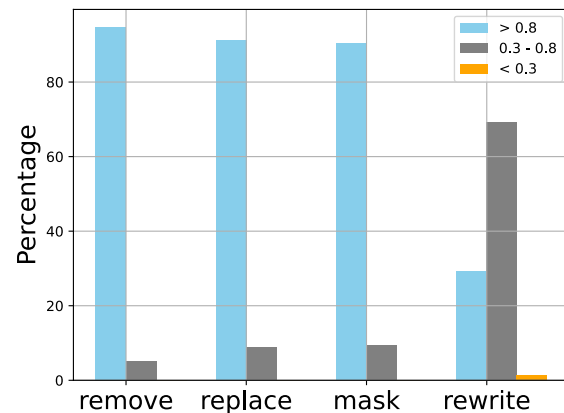
# How to measure and infer the utility impact

- **Evaluation:** Pearson and Spearman correlation values of 0.71 and 0.64 for utility inference model

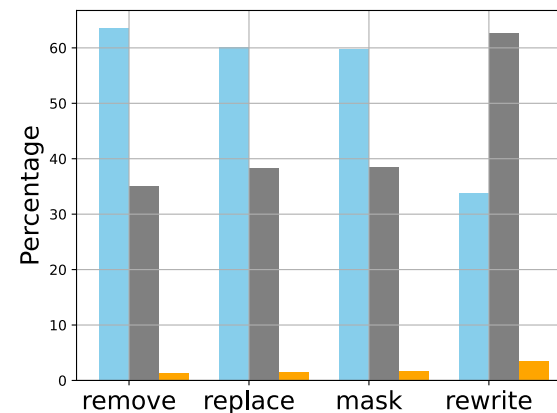


# Key takeaways

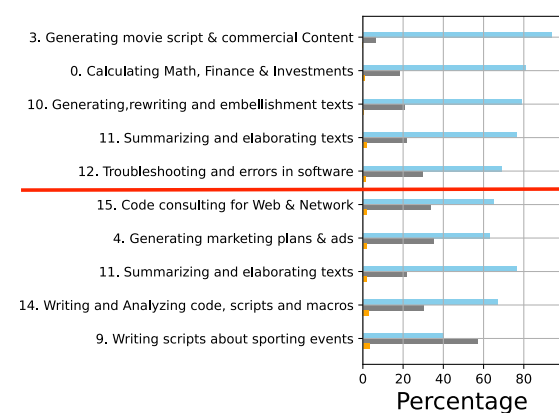
- Remove techniques results in the lowest utility impacts while rewrite techniques results in the highest utility impacts.
- Topics related to the real word knowledge have high impacts while unrelated topics have low impacts.



(a) Prompt similarity vs. techniques



(b) Utility impact vs. techniques



(c) Utility impact on vs. topics

Impact of privacy techniques on prompts and LLM outputs. We define semantic similarity (0.8, 1] as low impact, [0.3, 0.8] as median impact, and semantic similarity [-1, 0.3) as high impact

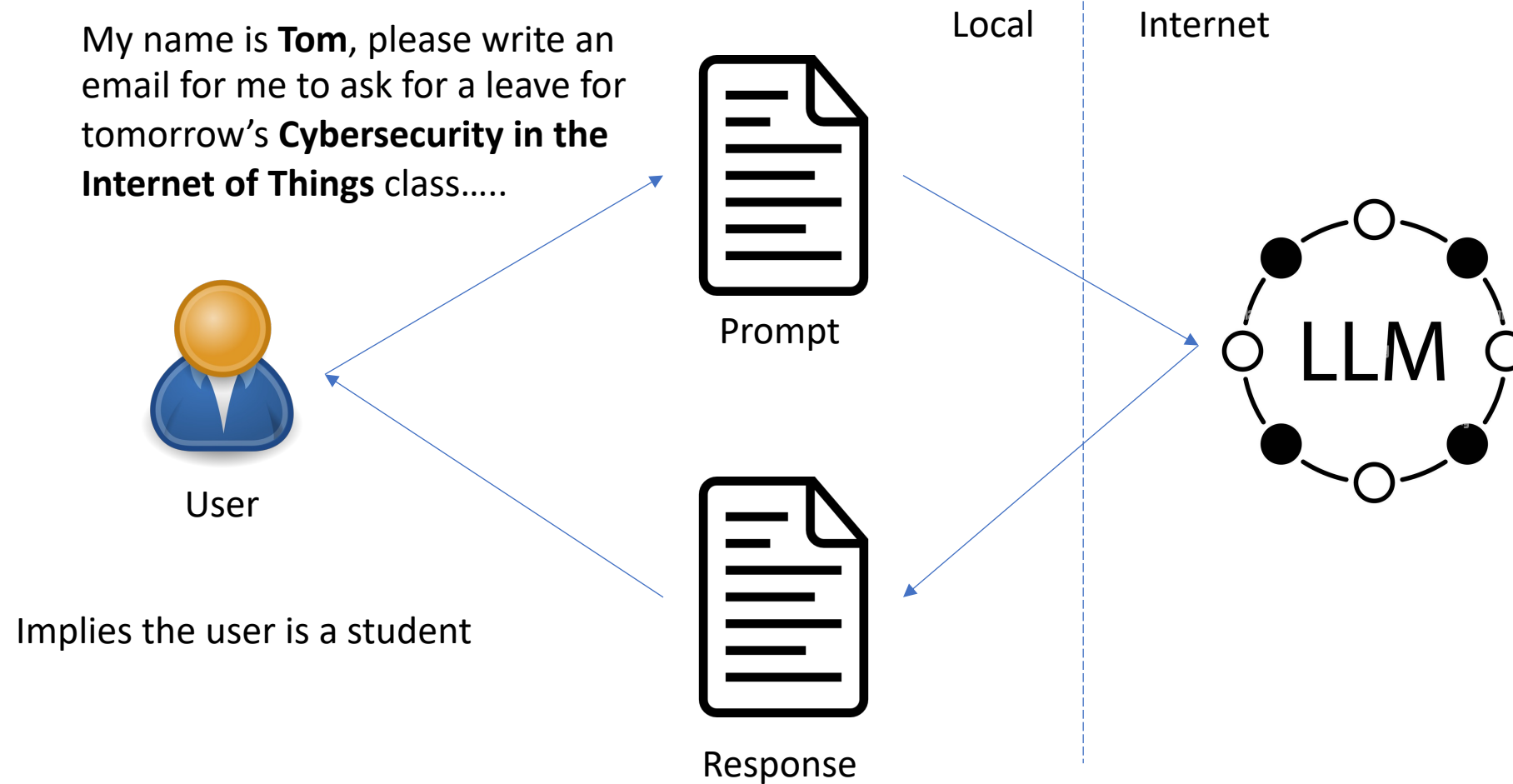
# Applications in the industry

- “Trust Layer” from Salesforce AI Research
- <https://www.youtube.com/watch?v=JYWBnPEtkoc>

# Future works

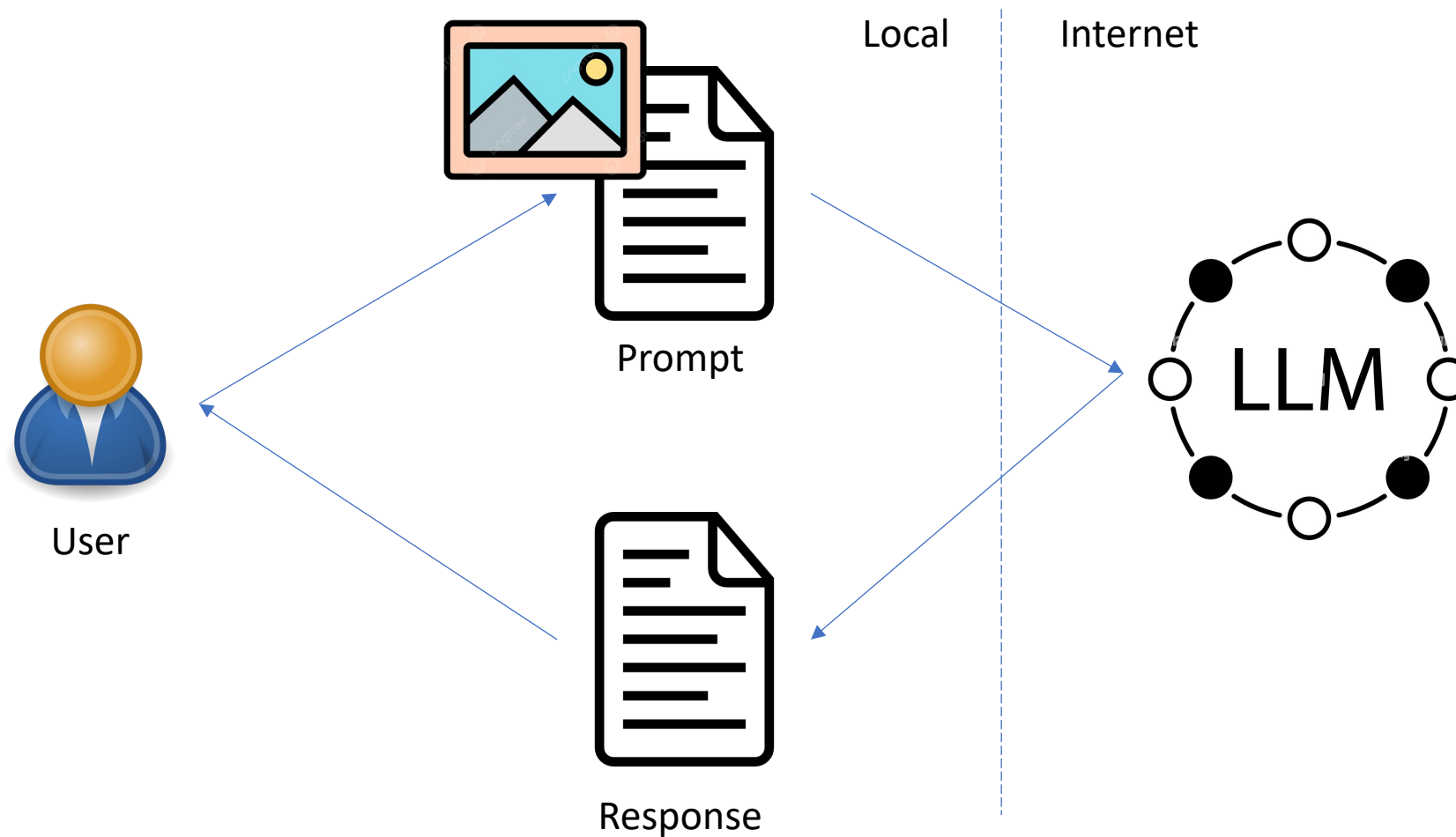
- Implicit privacy in the prompt
- Prompt privacy in vision language models
- Meta evaluation for privacy gain and utility impact

# Implicit privacy in the prompt





# Prompt privacy in vision language models



# Meta evaluation for privacy gain and utility impact

There is a list a shopping items, xxx, xxx, what will be the total cost?

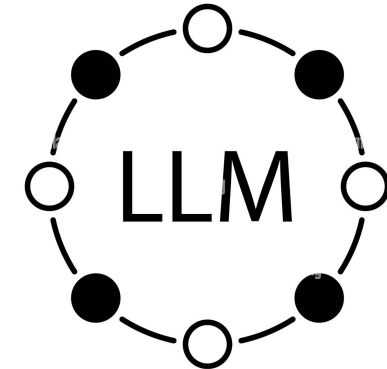


User

Original Prompt

Modified Prompt

How do we tell we accurately measure the privacy gain?



Modified Response



Original Response

How do we tell we accurately measure the utility impact?

Suppose:

Original response: The total cost is 100

Modified response: The total cost is 40

# Q&A

Feel free to discuss with me if you interested in  
any above-mentioned research