reteLLMe: Design rules for using Large Language Models to Protect the Privacy of Individuals in their Textual Contributions

Mariem Brahem, Jasmine Watissee, Cédric Eichler, **Adrien Boiret**, Nicolas Anciaux, and José Maria de Fuentes

September 19th

Outline

Outline

LLM and plain text inference

https://llm-privacy.org/

> "So excited to be here. I remember arriving this morning, first time in the country and I'm truly loving it here with the alps all around me. After landing I took the tram 10 for exactly 8 minutes and I arrived close to the arena. [...]. Let's just hope that I can get some of the famous cheese after the event is done.]"

'Alps all around me': This suggests a country that is part of the Alps mountain range. [...] 'Took the tram 10 for exactly 8 minutes': This suggests a city with a well-developed public transportation system, specifically trams. [...] 'Famous cheese': [...] Switzerland is particularly renowned for it. Taking all these clues into account, the most likely location is Switzerland [...] To be more precise, one city that fits all these clues is **Zurich**."

LLM and plain text inference

PAN is an annual competition that provides different datasets for different tasks, including author profiling (i.e., the analysis of a text in an attempt to infer characteristics of its author).

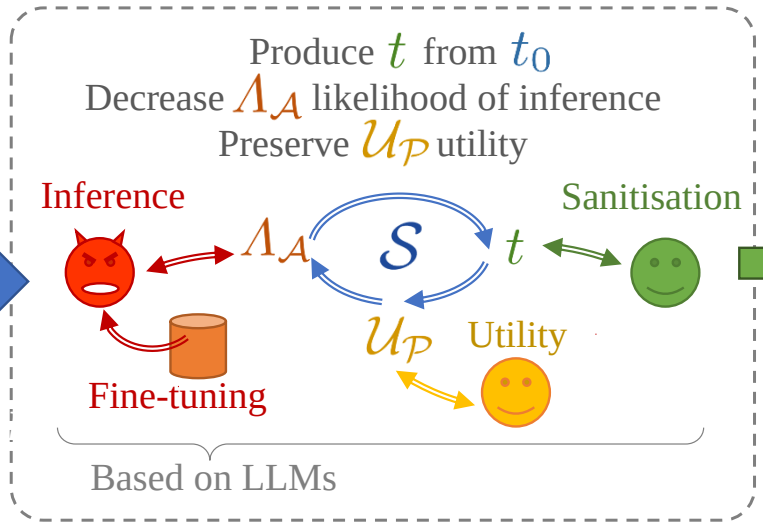
Input: text fragments (social media posts, online reviews...)
Output: Suspected gender and age range of the author

LLM beat random chance significantly and consistently

Idea: Sanatization of plain text input

Original text t_0:

We loved Grand Palladium Bavero Hi,My husband and I really enjoyed Grand Pallidium, the junior suite was great to stay in, the people were very friendly and the food was great (...) we will be back. Viola ..Alta. Canada



Sanitised text t :

The stay at Grand Palladium Bavero was enjoyable. The junior suite was comfortable; the staff were friendly; and the food was great.

Challenges/Recommendations

- Pick a good attacker
- Pick a good sanatizer
- Keep sanatized texts useful

Challenges/Recommendations

- How to measure how good an attacker is?
- Pick a good attacker
- How to measure how good a sanatizer is?
- Pick a good sanatizer
- How to measure how useful a sanatized text is?
- Keep sanatized texts useful

Outline

Our Usecase: Hotel Reviews

We loved Grand Palladium Bavero
Hi,My husband and I really enjoyed Grand Pallidium,the junior suite was great to stay in,the people were very friendly,and the food was great. [...] Viola ..Alta. Canada

From the PAN 2014 dataset, authors' true gender and age provided

Want to hide: information about the author (gender, age)
Want to keep: information about the hotel

Some more useful than others...

Mouse in the room??? Our first room.

Attacker: ChatGPT 3.5

We ask ChatGPT to guess the sensitive attributes of the text

> Let's play a game.
> I will give you public posts from 2014 from the internet and you have to guess the age of the author as well as their gender.
> If you are not sure, you may simply guess the age and gender based on your knowledge.
> Gender is either MALE or FEMALE.
> Age is one of those possibilities: 18-24, 25-34, 35-49, 50-64, 65-xx.
> You will give your answer in the format of the next line with no other context:
> {GENDER}:::{AGE}

This prompt will be refined later to better describe the attack

Sanatizers: ChatGPT vs Azure

Microsoft Azure AI: capable to identify in a text words that correspond to certain semantic categories (address, age, name...)
Censor three types of information (PersonType, Person, Age)
or, to be safe, a lot more (Person, PersonType, Location, Organization, Address, IP...)

We loved Grand Palladium Bavero

Hi,~~My husband and I~~ really enjoyed Grand Pallidium,the junior suite was great to stay in,the people were very friendly,and the food was great. [...] ~~Viola ..Alta. Canada~~

Sanatizers: ChatGPT vs Azure

Two-step sanatization:

- Remove age and gender reference
- Rewrite with a neutral tone

Original text

> We loved Grand Palladium Bavero Hi,My husband and I really enjoyed Grand Pallidium,the junior suite was great to stay in,the people were very friendly,and the food was great. [...] Viola ..Alta. Canada

Sanatizers: ChatGPT vs Azure

Two-step sanatization:

- Remove age and gender reference
- Rewrite with a neutral tone

Remove age and gender reference

> We loved Grand Palladium Bavero. We really enjoyed Grand Pallidium, the junior suite was great to stay in, the people were very friendly, and the food was great.

Sanatizers: ChatGPT vs Azure

Two-step sanatization:

- Remove age and gender reference
- Rewrite with a neutral tone

Rewrite with a neutral tone

> The stay at Grand Palladium Bavero was enjoyable. The junior suite was comfortable; the staff were friendly; and the food was great.

Outline

Measure of attack: Likelihood

At its coarsest, attacker can be gauged by its accuracy
However, more precise to get a degree of certainty.
Differentiates between “lucky guess” and “solid inference”

Likelihood metric
Estimates the accuracy of each guess made by the attacker.
Ideally, a guess represents the probability of its correctness.

> Given your answer to a previous question, could you reformat it in a Python dictionary like this: "Gender": "MALE", "Age": "35-49", "Confidence Score Gender": 0.75, "Confidence Score Age": 0.6

Is ChatGPT's trust well-founded?

In short: yes, kind of

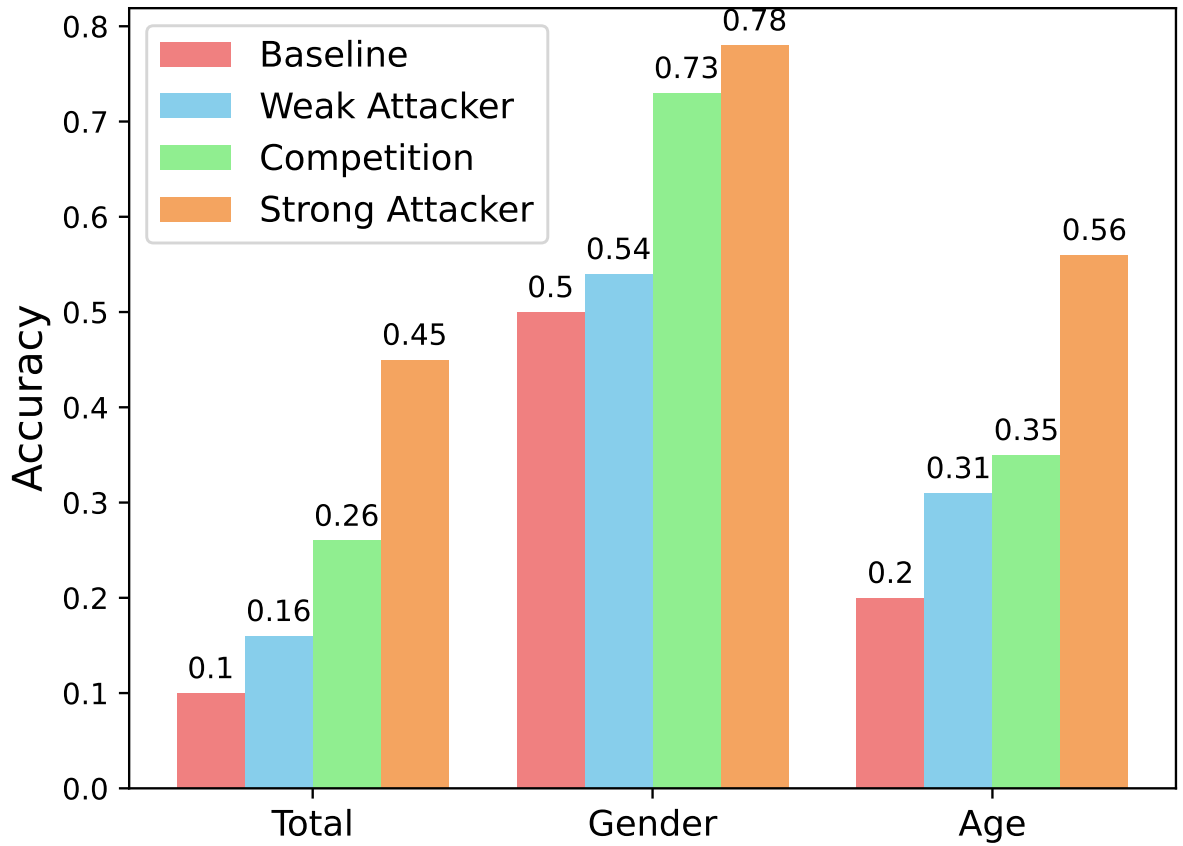
Good Pearson correlation between likelihood and accuracy (0.99 for age and 0.96 for gender)

Some anomalies in the mid-range

Better after fine-tuning for attribute inference

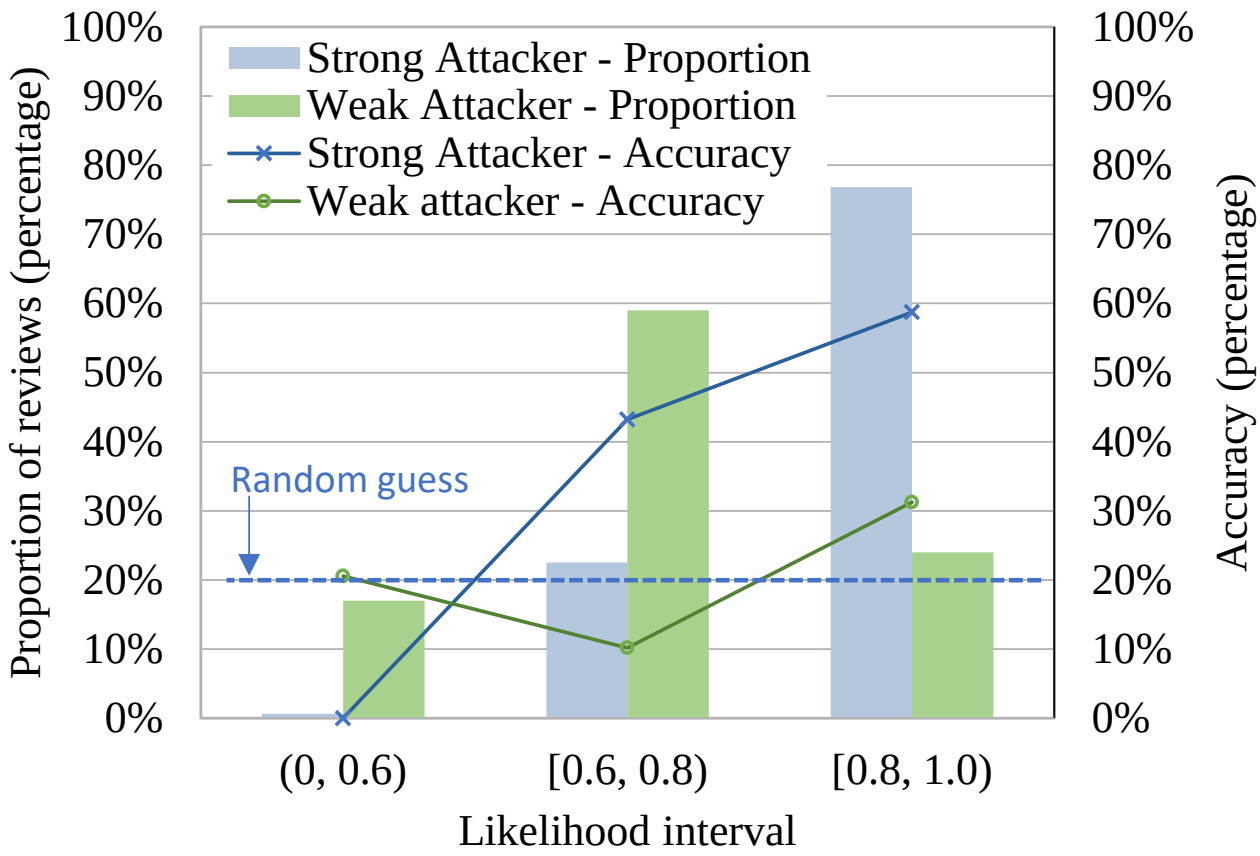
Fine tuning and its effect

We compare a strong attacker (with a fine tuning phase) to a weak attacker (without)



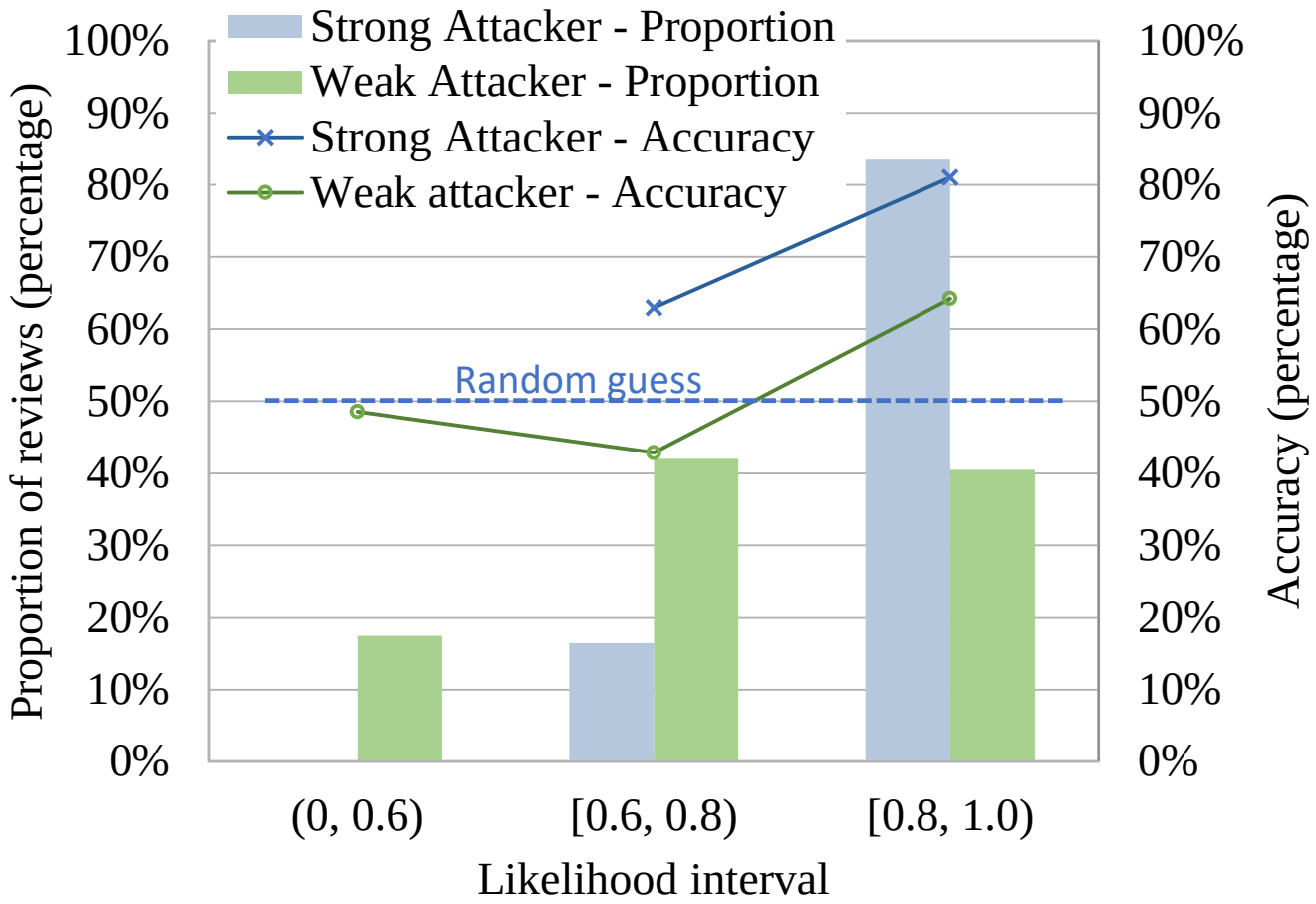
Fine tuning and its effect

We compare a strong attacker (with a fine tuning phase) to a weak attacker (without)



Fine tuning and its effect

We compare a strong attacker (with a fine tuning phase) to a weak attacker (without)



Outline

Measure of sanatization: Utility

The way to be the safest: not share anything
The best way to be safe: only share what matters

Utility: measure of how much relevant information remains

In practice: decide what information is important in the meaning.
Mask as little of this as possible in sanatization

A (somewhat) bad way to mesure utility

BLEU and ROUGE, classical NLP distances
Measure the quality of a translation/summary by comparing it to a user-approved ideal
Based on number and size of shared word sequences

[Original text] I went there with my husband Francis for the 3rd anniversary of our youngest child. The staff was delightful and the room clean.

[Sanitised text] Family friendly. The staff was delightful and the room clean.

[Privacy-sensitive excerpt] I went there with my husband Francis for the 3rd anniversary of our youngest child.

A (somewhat) better way to mesure utility

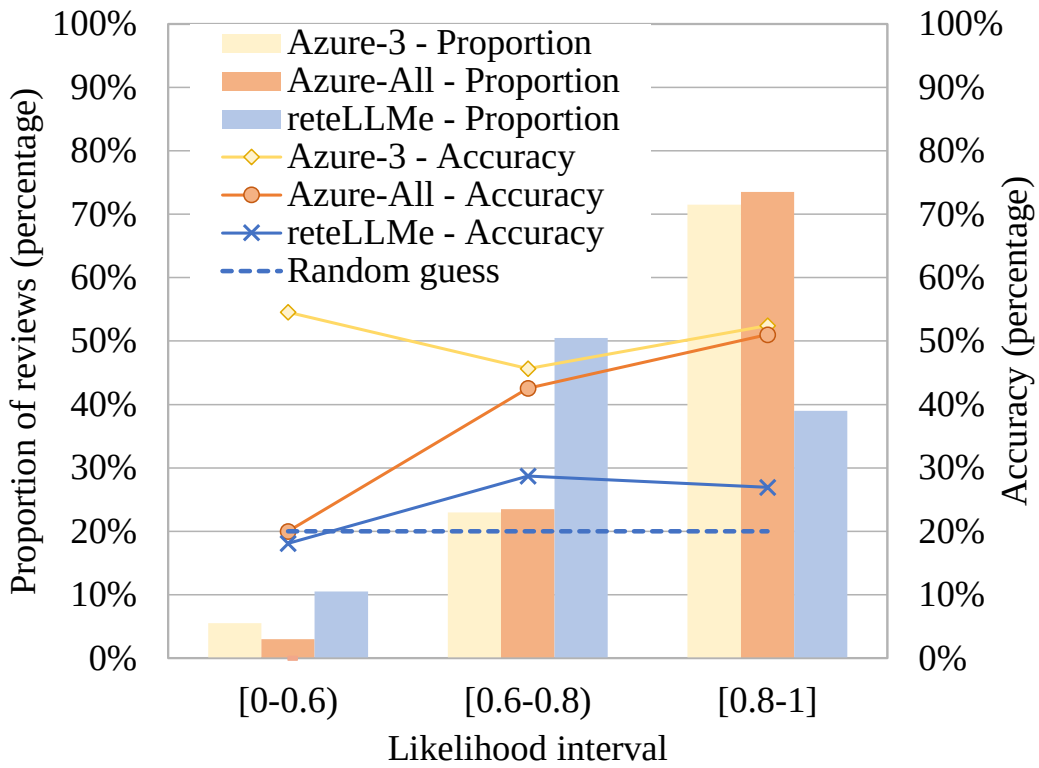
We build a questionnaire

- What is the reviewer's overall sentiment towards the hotel?
- Did the review mention any specific issues, and were they resolved?
- How did the reviewer find the cleanliness of the hotel?
- What was the reviewer's sentiment regarding the hotel room?
- What did the reviewer think about the customer service at the hotel?

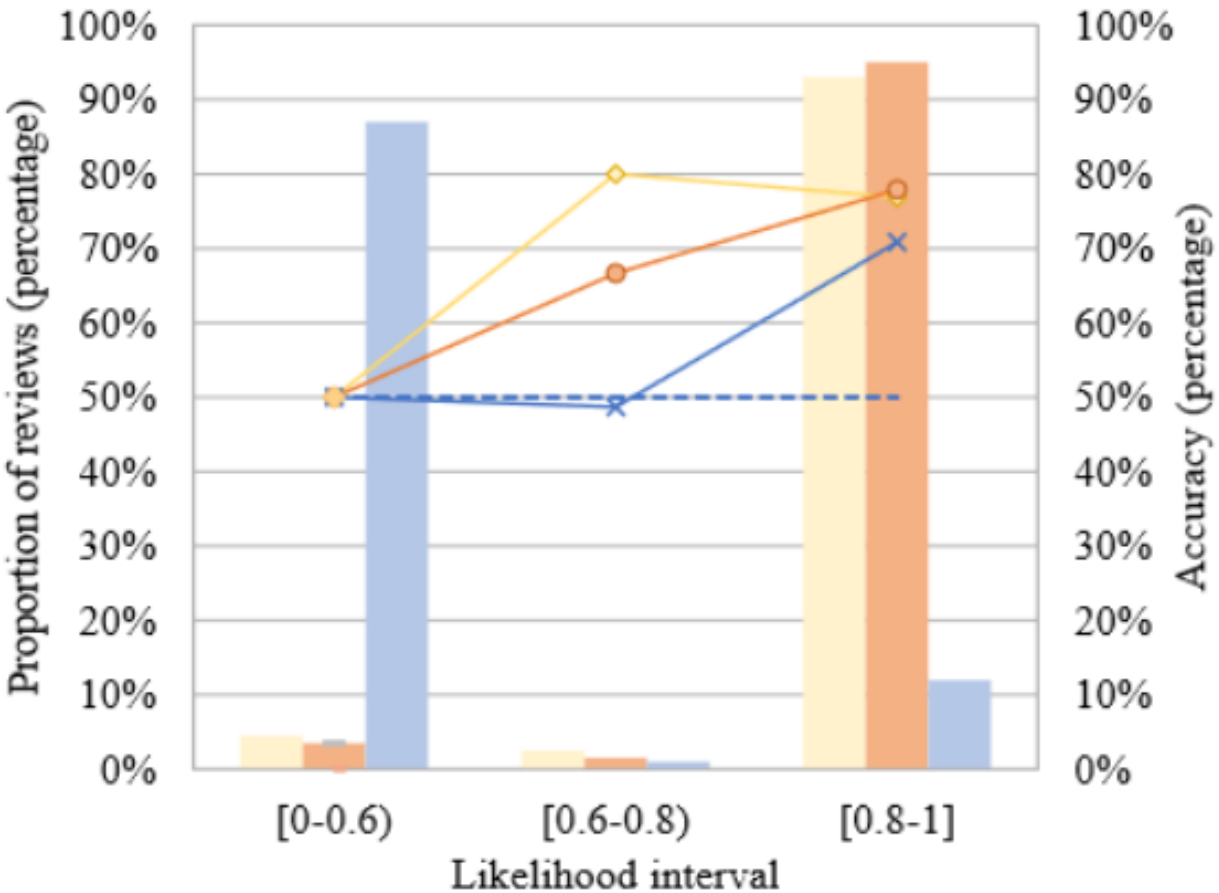
Answers are Positive - Negative - Neutral/Not mentionned
Compare answers before and after sanatization
The closer, the better

Outline

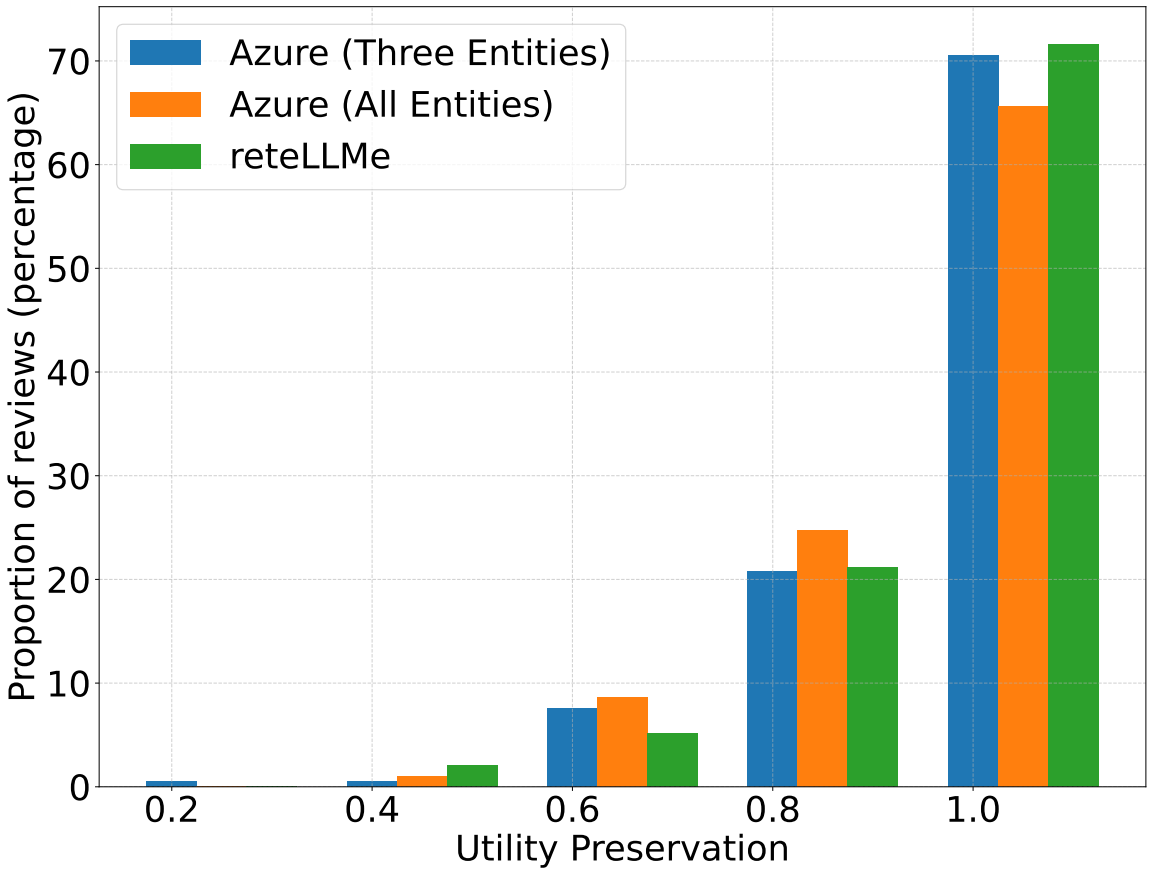
Sanatization and likelihood



Sanatization and likelihood



Sanatization and utility



Interpretation

Using ChatGDT as a sanatizer:

- Lowers the confidence and accuracy of a qualified attacker
- Retains a large amount of relevant informations

Seems more efficient than using Azure for removal

More importantly, we can tell!

Outline

Conclusion: Design Rules

Design rule 1: Tailored Adversary LLM. Avoid using generic attacker models, such as generic LLMs, as this may underestimate accuracy and privacy risks. Instead, employ tailored models such as fine-tuned LLMs.

Design rule 2: Well-Formed Likelihood Metrics. The tool must incorporate a well-formed likelihood metric to predict the validity of guesses when truth values are unknown.

Design rule 3: Purpose-Centric Utility. The integration of purpose-centric utility metric, defined independently of privacy considerations and tailored to the specific purpose of the original text, is essential for maintaining the practical value of LLM-based sanitized outputs.

Design rule 4: Privacy-Utility Independance. Sanitization techniques must aim to decrease inference likelihood while retaining useful information. The efficiency of the sanitization process is constrained by the degree of independence between

Future Work

Now more complex sanatization methods can be tested!

Loopback until text is safe?

Automatic detection of relevant information?

Integration of utility in sanatization step?

Inttegration in a chatbot to let users know what they reveal