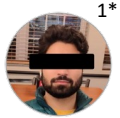
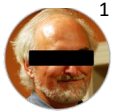
GenAIPABench:
A Benchmark for Generative AI-based Privacy Assistants

Privacy Enhancing Technologies Symposium, Bristol, UK and Online, 2024


1*
Aamir Hamid


1
Hemanth Reddy Samidi


1
Dr. Tim Finin


2
Dr. Primal Pappachan


1
Dr. Roberto Yus

1
UMBC

2
Portland State UNIVERSITY

**Most images in this presentation generated using Dall-E*

Privacy Policies and Data Regulations

Can tech decode privacy policies?

- Privacy policy length quadrupled since 2000, taking 304 hours a year to read [1].
- Average American adult reads at a 7th to 8th-grade level, but privacy policies require college-level understanding [2].
- 67% of people say they understand little to nothing about what companies do with their data [3].
- Research highlights increased collection and sharing of sensitive data and a lack of choice in data practices [4].

1. https://www.newscientist.com/article/2307117-privacy-policies-are-four-times-as-long-as-they-were-25-years-ago/
2. https://www.commonsense.org/education/articles/its-not-you-privacy-policies-are-difficult-to-read
3. https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy
4. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Related Work: Analysis of Privacy Policies

Manual Analysis:

- E.g., Mozilla Privacy Not Included [1].
- Too costly!
  - 68,160 minutes in 2022

Automatic Analysis (Using ML, DL, NLP, KG):

- E.g.,
  - Polisis [2]
  - PoliGraph [3]
  - ...
- Hard to understand users' questions and answer to them!

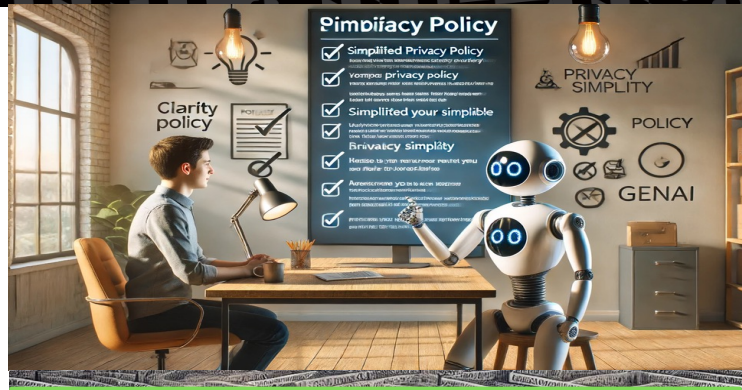
(new) LLM Analysis?

- E.g.,
  - PolicyGPT [4]
  - [5]

1. https://foundation.mozilla.org/en/privacynotincluded/
2. Harkous, Hamza, et al. "Polisis: Automated analysis and presentation of privacy policies using deep learning." 27th USENIX Security Symposium (USENIX Security 18). 2018.
3. Cui, Hao, et al. "PoliGraph: Automated privacy policy analysis using knowledge graphs." 32nd USENIX Security Symposium (USENIX Security 23). 2023.
4. Tang, Chenhao, et al. "Policygpt: Automated analysis of privacy policies with large language models." arXiv preprint arXiv:2309.10238 (2023).
5. Rodriguez, David, et al. "Large Language Models: A New Approach for Privacy Policy Analysis at Scale." arXiv preprint arXiv:2405.20900 (2024).

GenAI and LLM's

- Excels at processing large volumes of text swiftly.
- Enhances accessibility of complex documents.
- Understands context and intent behind user questions.
- Produces human readable answers.



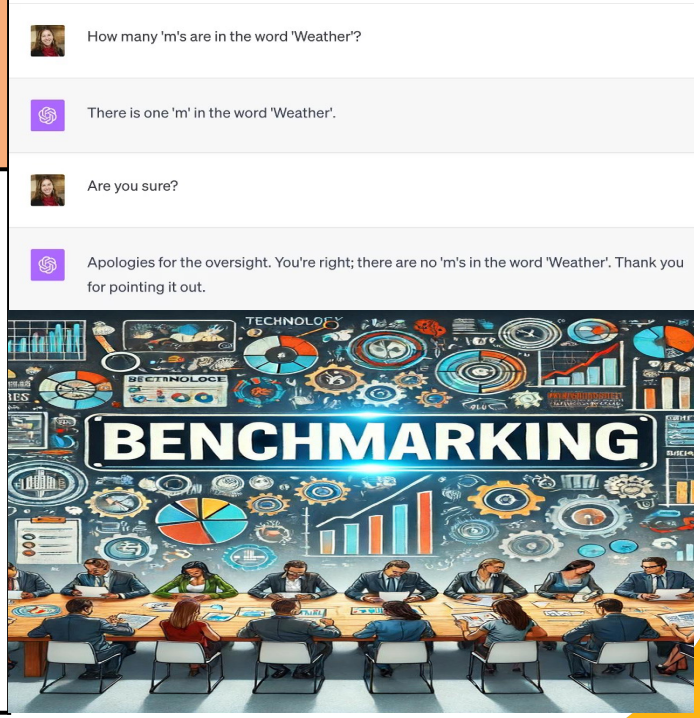
GenAI + Chatbot + Privacy Policy = GenAIPA?

Evaluation is Essential

GenAI can *"hallucinate"* (glitch), evaluation is important!

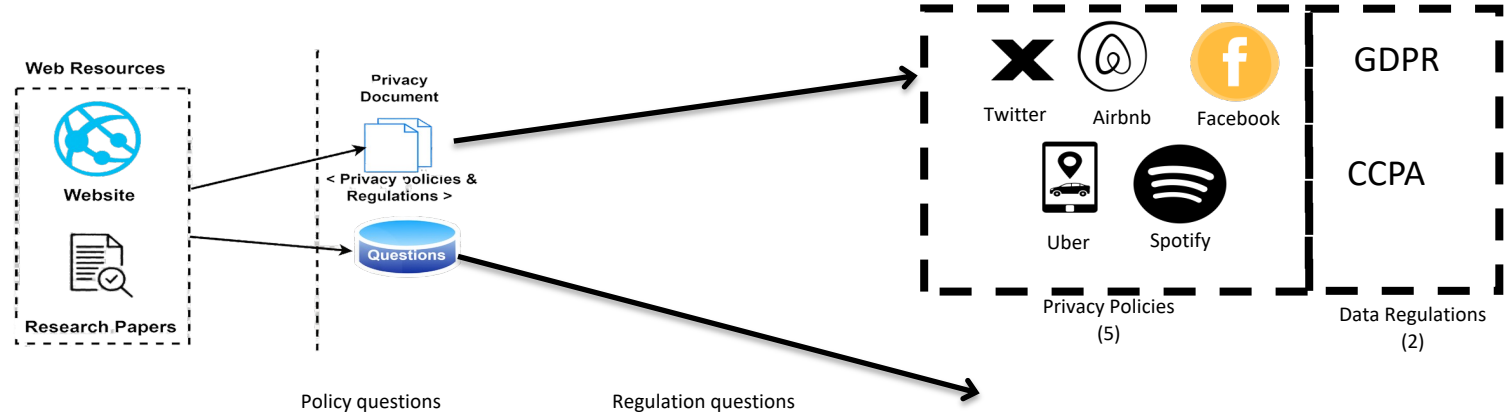
- *SAT Reading & Writing section*
  - GPT-4 → Score **710 / 800 (93rd percentile)** [1].
- *Bar exam*
  - GPT-4 → Score **298 / 400 (90th percentile)** [1].

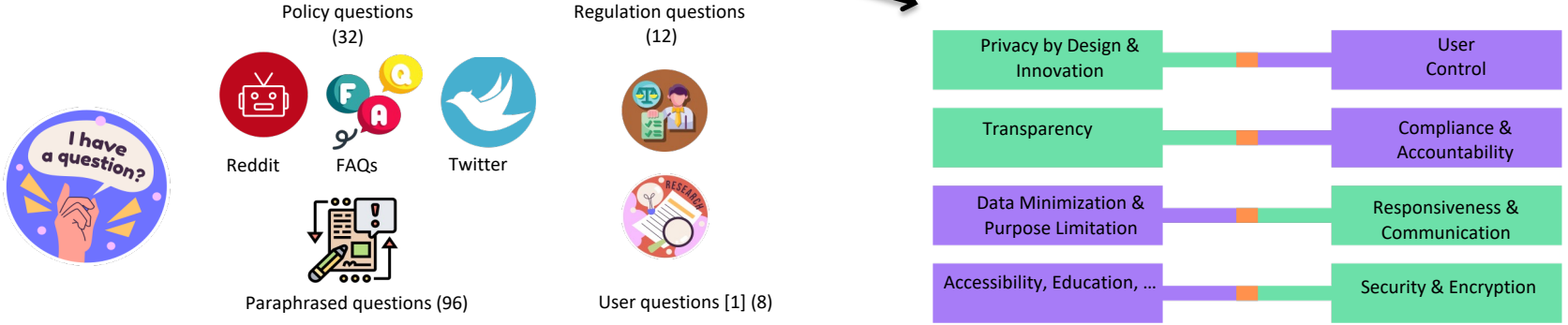
- **Where is our "privacy" exam for GenAI?**
  - **GenAIPABench**!



[1] https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1

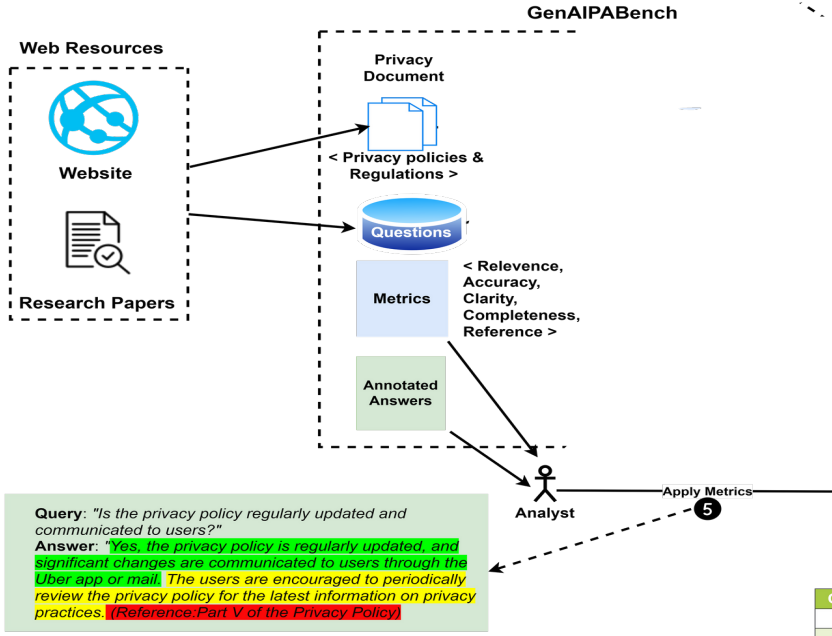
GenAIPABench: Privacy Documents & Questions

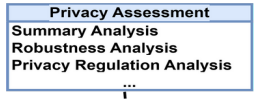




[1] Ravichander, A et al. "Question Answering for Privacy Policies: Combining Computational and Legal Perspectives". *ArXiv, abs/1911.00841*.

GenAIPABench: Metrics & Annotated Answers





Questions	M_{rel}	M_{acc}	M_{cla}	M_{com}	M_{ref}	Result
T_{f1}	1	1	1	1	1	
SE_{f2}	0.5	-1	1	-1	1	
RC_{f3}	-1	-1	-1	-1	-1	
...	...	...	...	...	...	

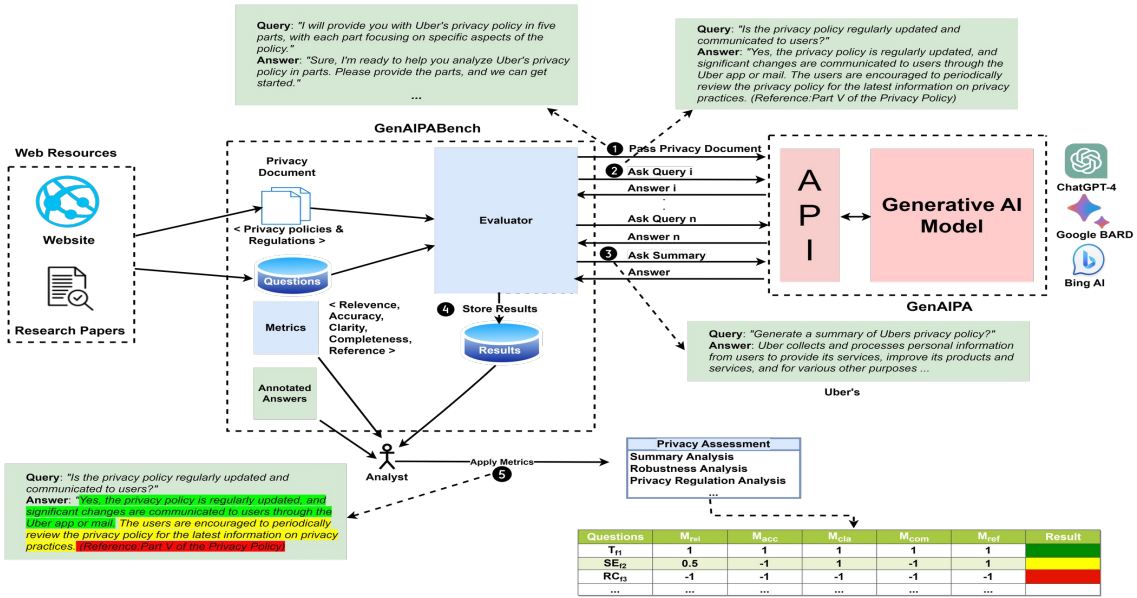
Metrics:

- Relevance
- Accuracy
- Completeness
- Clarity
- Reference

Annotated Answers:

- 200+ expert generated answers

GenAIPABench: **Evaluator**



Evaluator Functions:

- Automates Prompt Generation.
- Executes Prompts.

Types of Initialization Prompts:

- Without privacy document.
- With segmented portions of the document.
- Requesting summary first.

Using GenAIPABench

Analysis 1: Quality of Responses to Privacy Policy Questions

Analysis 2: Robustness through Paraphrased Questions

Analysis 3: Ability to Recall Learned Privacy Policy Knowledge

Analysis 4: Quality of Responses to Privacy Regulation Questions

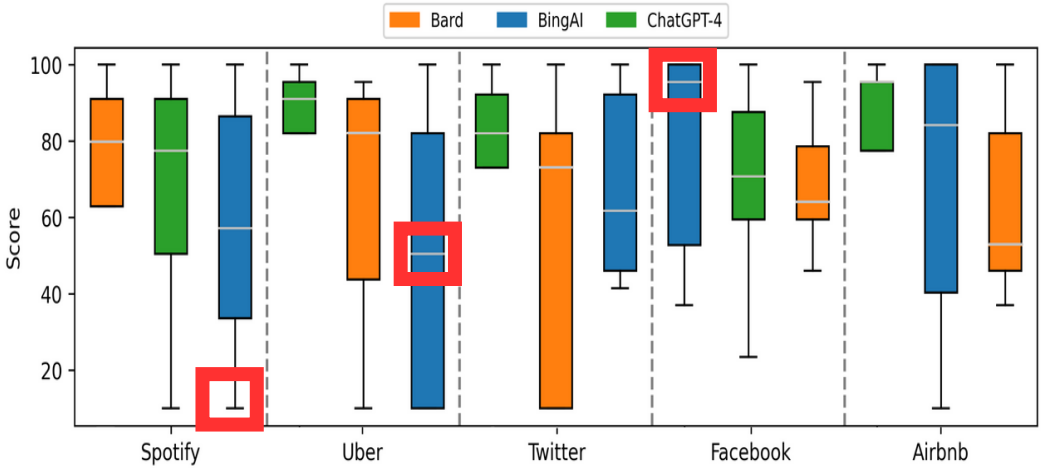
Analysis 5: Quality of Privacy Policy Summaries

GenAIPAs Used:

- ChatGPT-4
- Google Bard
- Bing AI

Experiments Date:
(July - Aug) 2023

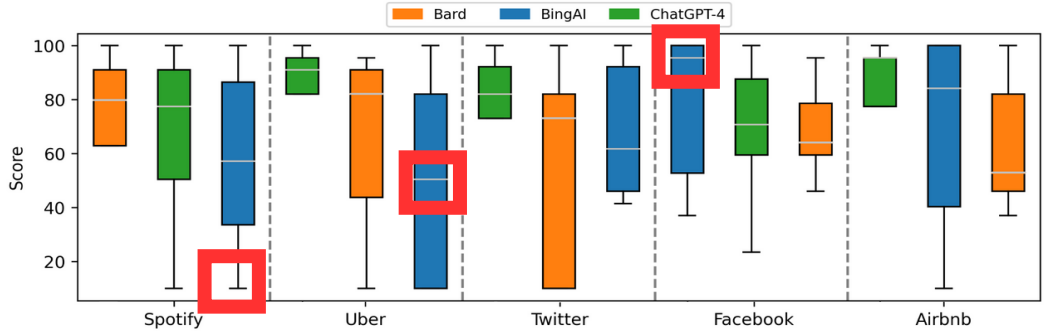
Quality of Responses to Privacy Policy Questions



Good performance when executing benchmark for multiple runs…

In some runs the performance is really bad! (less than 10/100)

Quality of Responses to Privacy Policy Questions



Good performance when executing benchmark for multiple runs…

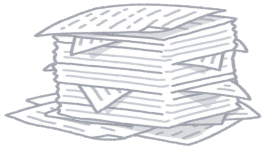
In some runs the performance is really bad! (less than 10/100)

Answers are clear and relevant but…
- Fails badly at referencing!
- Accuracy is a hit and miss…

(b) Average scores for all policies across metrics.

Discussions

GenAI systems struggled more with **compliance and accountability** topics.



Policy length, rather than reading level, significantly affected system performance.

Questions with **explicitly defined policy content** were easier for systems to handle.

Paraphrased questions reduced system performance, highlighting the need for user-friendly queries.

Systems scored higher on **privacy regulations** due to more online discussion compared to specific policies.

Conclusions

- **GenAI offers potential** for advanced privacy assistants (GenAIPAs) but needs rigorous evaluation.
- **GenAIPABench first benchmark for GenAIPAs on privacy policies and regulations.**
- Current GenAI show promise but **struggle with paraphrasing, referencing, and accuracy** (among others).
- **We plan to keep the benchmark updated!**

Thank You!

Artifacts are available :

GenAIPABench is open-source and available on GitHub

https://github.com/DAMSlabUMBC/GenAIPABench