GenAIPABench: A Benchmark for Generative AI-based Privacy Assistants

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

*Most images in this presentation generated using Dall-E
Privacy Policies and Data Regulations

- 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
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]

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.
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

Web Resources
- Website
- Research Papers

Privacy Document

Questions
- Privacy policies & Regulations

Policy questions (32)
- Reddit
- FAQs
- Twitter
- Paraphrased questions (96)

Regulation questions (12)
- User questions [1] (8)

GDPR
- Twitter
- Airbnb
- Facebook

CCPA
- Uber
- Spotify

Privacy Policies (5)
Data Regulations (2)

User Control

Compliance & Accountability

Responsiveness & Communication

Security & Encryption

Privacy by Design & Innovation

Transparency

Data Minimization & Purpose Limitation

Accessibility, Education, ...

GenAIPABench: Metrics & Annotated Answers

**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!**

Artifacts are available:

GenAIPABench is open-source and available on GitHub

https://github.com/DAMSLabUMBC/GenAIPABench