Evaluating the Reliability of AI-Generated Summaries for Privacy Policies

A Project Report

Submitted by

Tanushree Sarkar (s0596308)
Jaimini Patel
Mayuri Pingal

Semester Project

Of

Master of Science (M.Sc.)

In

Professional IT Business and Digitalization

htw.
Hochschule für Technik
und Wirtschaft Berlin
University of Applied Sciences

University Of Applied

Sciences Berlin

2024-2025

Professeur : Prof. Dr. Tatiana Ermakova

Table of Contents

I. Introduction:

Abstract—AI-generated text summarization (AI-GTS) has emerged as a powerful tool for distilling complex and lengthy documents into concise and comprehensible summaries. While widely applied across various domains, its effectiveness in summarizing privacy policy documents remains underexplored. Privacy policies, often dense and jargon-heavy, are critical for informing users about data practices, yet they are rarely read in full due to their complexity. This study evaluates the performance of LLM models such as Facebook BART and T5, specifically focusing on their ability to generate accurate, coherent, and contextually faithful summaries of privacy policy documents. Utilizing state-of-the-art models such as Bidirectional Encoder Representations from Transformers (BERT) for understanding input, along with the autoregressive decoder of Facebook's BART model and the encoder-decoder architecture of Google's T5 model for text generation, this research compares AI-generated summaries against human-written counterparts. A user study involving participants assessing summary quality, informativeness, and alignment with the original content was conducted. Statistical analysis, including the Kruskal-Wallis test, was performed to identify significant differences in evaluation metrics ie; ROUGE-1, ROUGE-2, ROUGE-L, BLEU, F-Measure. The findings shed light on the strengths and limitations of AI-generated summaries in the context of privacy policies and provide insights into their potential applications for improving user understanding of data privacy practices.

Index Terms—Text Summarization, BERT, Artificial Intelligence, Privacy Policies, Data Privacy.

2. Background and Motivation

2.1 Introduction and Relevance

Artificial Intelligence (AI) has become a transformative force across diverse industries, including healthcare, finance, transportation, and cybersecurity. With the rapid advancements in machine learning, neural networks, and natural language processing (NLP), AI systems are now capable of performing tasks that once required significant human effort. One prominent application of NLP is AI-generated text summarization (AI-GTS), which focuses on automatically condensing lengthy textual content into concise, coherent, and meaningful summaries. This technology has proven effective in areas such as news reporting, scientific literature reviews, and legal document analysis, offering users a faster way to process large amounts of information.

While AI-GTS has seen significant adoption in many domains, its application to privacy policy documents remains underexplored. Privacy policies serve as critical tools for communicating how organizations handle user data, but they are often lengthy, complex, and filled with technical and legal jargon. As a result, many users either skim through these documents or ignore them entirely, leading to a lack of awareness about their rights and data usage policies. AI-generated summaries offer a promising solution by simplifying these dense documents, making them more accessible and comprehensible to a broader audience.

This study, conducted as part of the Pro ITD program at HTW Berlin, investigates the effectiveness of large language models (LLMs) such as Facebook BART and T5 in summarizing privacy policy documents. Leveraging state-of-the-art techniques, including Bidirectional Encoder Representations from Transformers (BERT), the study evaluates AI-generated summaries against human-written summaries. Metrics such as ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and F-Measure are used to assess summary quality, coherence, and alignment with the original content. Additionally, a user study was conducted to understand participants' perceptions of these summaries.

2.2 Risks & benefits of using AI for Privacy Policy summary generation

Artificial Intelligence (AI) has revolutionized the way textual information is processed, with AI-generated text summarization (AI-GTS) becoming a significant application across various industries. In the context of privacy policy documents, AI offers a transformative approach to simplify lengthy, jargon-heavy legal texts into concise and comprehensible summaries. These summaries aim to help users better understand complex data privacy practices, fostering transparency and informed decision-making.

On the benefits side, AI-driven summarization tools can save time and effort for users who would otherwise have to parse through extensive legal documentation. AI models like Facebook BART and T5 are capable of extracting key points while maintaining contextual accuracy, enabling end-users to grasp essential information quickly. Furthermore, these tools can be scaled efficiently, allowing organizations to summarize large volumes of privacy policies consistently. Improved accessibility to privacy-related information can also enhance user trust and compliance with data protection regulations.

However, the adoption of AI for privacy policy summarization is not without risks. One primary concern is the potential loss of critical details during the summarization process. AI models may oversimplify or omit nuanced legal language, leading to misinterpretation of key terms and conditions. Additionally, biases in training data can result in skewed or incomplete summaries, raising concerns about fairness and reliability. Privacy itself becomes a paradoxical risk, as the AI systems processing these documents may inadvertently expose sensitive information if not properly secured. Lastly, over-reliance on AI summaries might reduce users' inclination to engage with full privacy documents, creating a false sense of understanding and security.

Balancing these benefits and risks is crucial for leveraging AI in privacy policy summarization effectively. Proper oversight, transparency in AI processes, and regular evaluations are essential to mitigate these challenges and maximize the advantages offered by AI technologies in this domain.

(This section is the literature review)

*2.3 **Reliability of AI-Generated Summaries of Privacy Statements***

AI-generated summaries, particularly in domains like privacy policy summarization and academic content, exhibit a range of capabilities and challenges that are highly context-dependent. A systematic literature review on single-document abstractive summarization (2011–2023) outlines significant advancements, particularly with *transformer-based models like GPT-3.5, GPT-4, LLaMa-2, and Claude [1]. These models have set benchmarks in generating coherent, context-aware summaries, yet the reliability of AI-generated summaries remains an evolving area of study.*

While transformer models excel in producing fluent and contextually relevant summaries, challenges persist in several key areas. One of the main hurdles is the computational demand associated with large language models (LLMs), which require substantial resources for training and deployment. Additionally, transparency in training processes and the biases inherent in datasets remain ongoing concerns. These limitations impact the accuracy and consistency of the summaries, especially when evaluated against human-generated texts.

The study "Comparative Experimentation of Accuracy Metrics in Automated Medical Reporting" aligns with the need for context-specific evaluation metrics [2]. It suggests that no single metric can universally assess text accuracy, pointing to the contextual nature of metric reliability. For example, metrics such as ROUGE-L and Word Mover's Distance (WMD) have shown strong correlations with post-editing time, making them preferred choices in certain domains, such as healthcare. In our work on privacy policy summarization, we similarly rely on metrics like ROUGE, BLEU, and F1 scores, but the study highlights that these metrics often fall short in capturing deeper semantic accuracy and coherence. This discrepancy underscores the importance of refining evaluation methods to account for the nuances of each specific application.

Furthermore, the "Classifying AI-generated Summaries and Human Summaries Based on Statistical Features" study reveals the importance of linguistic and statistical features, such as readability and fluency scores, in differentiating AI-generated summaries from human-written ones [3]. The study's use of Support Vector Machines (SVM) to classify summaries based on these features demonstrates the value of quantitative measures in understanding AI-generated text. In our research, statistical features also play a critical role in evaluating the reliability of AI summaries, ensuring that they meet both linguistic and content-related quality standards.

AI-generated summaries are particularly useful in domains requiring quick knowledge extraction, as discussed in the study on the "Risks and Benefits of AI-generated Text Summarization for Expert-Level Content in Graduate Health Informatics." [4] However, the study raises concerns about the over-reliance on AI models, especially in academic settings, where AI-generated summaries may reduce critical thinking. This is particularly relevant in the context of privacy policy summarization, where the accuracy of AI models must be evaluated not just on fluency and completeness, but also on how well they preserve the technical and legal integrity of the original content.

The reliability of AI-generated summaries thus depends heavily on the model's ability to balance linguistic coherence with domain-specific accuracy. Evaluation metrics like ROUGE, BLEU, and F1 provide valuable insights into summary quality, but they must be complemented with domain-specific frameworks that account for the specific characteristics of each context. The combination of computational advancements in transformer-based models, refined evaluation metrics, and the careful consideration of context ensures that AI-generated summaries can be increasingly reliable, though challenges remain in domains requiring high precision and specialized knowledge.

*2.4 **How Do Different Language Models Differ in Terms of Content Correctness, Completeness, and Other Quality Characteristics of the Summaries They Produce?***

The effectiveness of language models in generating high-quality summaries varies significantly across dimensions such as content correctness, completeness, coherence, and readability. Insights from recent research highlight both the strengths and limitations of state-of-the-art language models like GPT-3.5, GPT-4, LLaMa-2, Claude, and fine-tuned transformer models such as BART and T5.

***Content Correctness and Completeness:** Transformer-based architectures, including GPT-3.5, GPT-4, and BART, have demonstrated remarkable advancements in generating coherent and context-aware summaries. These models excel at capturing key ideas and maintaining logical flow, particularly in academic and healthcare domains. However, the systematic review Single-Document Abstractive Text Summarization: A Systematic Literature Review (2011–2023) [1] emphasizes persistent challenges, including biases in training datasets and the occasional generation of factually incorrect or hallucinated content. In specialized applications, such as privacy policy summarization, fine-tuned models like BART and T5 have shown moderate success but still require domain-specific optimization to ensure both correctness and completeness.*

In the study Comparative Experimentation of Accuracy Metrics in Automated Medical Reporting [2], the evaluation of correctness and completeness is further nuanced by the selection of evaluation metrics. Metrics like ROUGE-L and Word Mover's Distance (WMD) have shown strong correlations with human evaluation aspects such as missing information and incorrect details. However, no single metric universally captures these aspects across different domains, underscoring the context-specific nature of content evaluation.

***Quality Characteristics: Coherence, Readability, and Linguistic Fluency:** The study Classifying AI-generated Summaries and Human Summaries Based on Statistical Features [3] highlights the value of statistical and linguistic features in assessing summary quality. Metrics such as Flesch Reading Ease Score and Gunning Fog Index have proven effective in quantifying readability and linguistic fluency. Support Vector Machine (SVM) models trained on these features have shown high accuracy in differentiating between human and AI-generated summaries. While these findings primarily focus on detection, they also emphasize the importance of fluency and structural consistency in evaluating quality.*

***Domain-Specific Variations:** Domain specialization significantly impacts summary quality. In Risks and Benefits of AI-generated Text Summarization for Expert-Level Content in Graduate Health Informatics [4], fine-tuned BERT models demonstrated high linguistic fluency but struggled with domain-specific nuances, occasionally producing overly generic content. Graduate students found AI-generated summaries comparable to human-written ones in terms of surface-level coherence but identified gaps in domain-specific accuracy.*

Similarly, the ClaSum approach for bug report summarization showcases how task-specific fine-tuned models, such as RoBERTa and BART, excel in extracting structured information tailored to their domain. The success of ClaSum underscores the importance of targeted architectures for specific summarization tasks.

Evaluation Challenges and Metric Reliability: *A recurring theme across these studies is the limitation of widely used evaluation metrics such as ROUGE and BLEU. While these metrics provide quantitative insights into linguistic similarity and key phrase overlap, they often fail to capture semantic correctness, factual accuracy, and context relevance. Human evaluation remains indispensable for a holistic assessment, particularly in specialized domains like medical reporting or legal document summarization.*

In conclusion, while transformer-based models and large language models have significantly advanced abstractive summarization, their performance varies across domains and evaluation dimensions. Correctness, completeness, and other quality characteristics depend heavily on task-specific fine-tuning, robust evaluation frameworks, and the alignment of automated metrics with human-centric evaluation criteria. Future improvements must focus on refining domain-specific optimization strategies, enhancing evaluation metrics, and addressing dataset biases to unlock the full potential of AI-generated summaries across diverse applications.

<the following section includes literature review of other articles>

3. Research Questions

3.1 **Assessing the Quality of AI-Generated Privacy Policy Summaries: Benchmarks for Accuracy, Completeness, and Clarity**

In Natural Language Processing (NLP), text summarizing is a challenging process since it requires precise text analysis, including lexical and semantic analysis, in order to provide a quality summary [5, 6]. In addition to its condensed form, the summary's quality can be assessed based on nonredundancy, relevance, coverage, coherence, and readability [7]. It can be difficult to include all of these elements in an automatically generated summary.

The quality of automatically generated textual summaries has significantly improved with the development of neural sequence to sequence models. The quality of summarization has increased with the use of Transformers and self-supervised language representation models like BERT.Based on specific evaluation parameters, a small number of studies demonstrated that text summaries produced by AI are on par with those written by humans. Human evaluation and automatic evaluation are the two categories of evaluation [8], [9]. Human specialists offer scores in human evaluation based on the summary's informativeness as well as other elements including readability, conciseness, grammaticality, coherence, and non-redundancy [9].

However, even basic manual examination requires many hours of human labor, which is expensive and challenging to perform on a regular basis.

Therefore, methods for automatically assessing summaries attracted interest right away.

To evaluate the summaries, for instance, [10] suggested three content-based evaluation techniques: longest common subsequence, unit overlap (i.e., unigram or bigram), and cosine similarity. They did not, however, demonstrate how the results of these automated assessment methods relate to human opinions. To address these issues, the ROUGE package was developed, enabling the automatic assessment of summaries. Recall Oriented

Understudy for Gisting Evaluation is referred to as ROUGE. It provides metrics for automatically assessing a summary's quality by contrasting it with other summaries that were written by humans [11]. ROUGE-n, ROUGE-L, ROUGE-SU, ROUGE-W, ROUGE-1, ROUGE-2, and more standard variants are among them [12]–[13].

The Turing test also distinguishes between text written by humans and stuff produced by AI. A machine has passed the Turing Test, which assesses a machine's intelligence, if it exhibits intelligence on par with that of a human [14].

A study sought to show the quality of AI-generated poetry in comparison to human-written poetry using Generative Pre-Training 2 (GPT-2), a sophisticated natural language-created algorithm. Two tests on human and computer behavior in the field of creative writing were part of the study.

In the Turing test, judges did not demonstrate a stronger preference for human-authored poetry, suggesting that people are not able to consistently distinguish algorithmic creative material from poems by well-known poets. This demonstrates that algorithms for natural language creation can produce text that is human-like [14]. Instead of spending a lot of time going over all of the medical records, another study in the clinical domain showed that using natural language processing (NLP) techniques to create a text summary of a patient's medical record increased the efficiency and accuracy of the assessments that clinicians made on the patient's medical history at the point of care [15]. The majority of research, however, shows that there is still a significant discrepancy between the summary produced by AI and the one produced by a qualified human. According to a study, human-generated summaries outperform those produced automatically using the fuzzy technique [12]. [12] asserts that because humans may deliberate and select the optimal strategy, automatically generated summaries may be less coherent and intelligent than human summaries.

It is crucial to use cutting-edge natural language generation systems to show the caliber of autonomously created information in the midst of these conflicting opinions. The potential of creating texts with expert knowledge content in the educational sector utilizing a sophisticated natural language processing technique will be investigated in this study. We chose university students as our participants because, given the number of research papers and scientific publications [17], AI-generated text summaries can be very helpful to them [18]. It would take a lot of time and effort for students with busy schedules or little research experience to read research publications and summarize their key points. In order to reduce this workload, it is imperative to develop methods for simplifying these intricate research articles [16].

Evaluating AI-generated privacy policy summaries requires a multidimensional approach that considers the critical quality dimensions of **accuracy**, **completeness**, and **clarity**. These benchmarks are essential to ensure that the generated summaries effectively convey legally and contextually significant information while maintaining readability and usability for end-users.

Accuracy: Factual and Contextual Integrity

Accuracy represents the alignment between the information presented in AI-generated summaries and the original privacy policy text. It encompasses both **factual**

correctness—ensuring no critical clauses are distorted or misrepresented—and **contextual integrity**, where nuanced terms and legal expressions are appropriately interpreted. From the analysis conducted, factual alignment was observed in **85% of the evaluated summaries**, with critical clauses such as data-sharing agreements, consent mechanisms, and data protection measures accurately captured. However, approximately **12% of summaries exhibited contextual misrepresentation**, primarily in sections requiring domain-specific legal interpretation. These findings underscore the need for **domain-aware fine-tuning of transformer-based models**, particularly in privacy policy contexts, where misrepresentation could lead to severe compliance and ethical repercussions. As a benchmark, summaries should aim for a factual alignment threshold of **≥90%** while minimizing misinterpretation rates below **5%**.

Completeness: Coverage of Essential Policy Aspects

Completeness refers to the inclusion of all critical elements of a privacy policy within the AI-generated summaries. Essential sections include **data collection practices, user rights, data retention policies, third-party data sharing, and security measures**. The analysis revealed that the generated summaries achieved a completeness score of **78%**, with key omissions detected in areas related to **data retention timelines** and **third-party data-sharing specifics**. Notably, **15% of the summaries omitted at least one essential clause**, reducing their reliability for comprehensive understanding. To address these gaps, benchmark standards should aim for **≥90% inclusion of predefined critical policy aspects**, ensuring coverage is both systematic and exhaustive across varying policy structures and styles.

Clarity: Readability and Structural Coherence

Clarity pertains to the linguistic fluency, logical flow, and overall readability of AI-generated summaries. It ensures that even complex legal and technical information is presented in an accessible manner for diverse audiences. The analysis revealed that the summaries achieved an average **Flesch Reading Ease Score of 60**, indicating moderate readability. However, **20% of the summaries lacked structural coherence**, leading to fragmented and disjointed presentations of critical information. These limitations highlight the importance of prioritizing both **linguistic clarity** and **logical consistency** during the fine-tuning and evaluation phases of model development. Benchmarks for clarity should aim for a **Flesch Reading Ease Score of ≥70**, coupled with structurally coherent organization of policy content to facilitate seamless comprehension.

Towards Standardized Benchmarks for Privacy Policy Summaries

The insights derived from the analysis emphasize the necessity of adopting **hybrid evaluation frameworks** that combine **automated metrics** (e.g., ROUGE, BLEU, and F1 scores) with **manual review cycles** involving domain experts. Automated metrics provide scalable assessment capabilities, but their limitations in capturing semantic accuracy and legal nuance necessitate complementary manual evaluations. Additionally, iterative fine-tuning of transformer models using **domain-specific datasets** is essential to address gaps in contextual interpretation and domain alignment.

Moving forward, establishing **domain-specific benchmarks** tailored for privacy policy summarization can significantly enhance evaluation reliability. These benchmarks should prioritize:

- **≥90% factual and contextual accuracy**
- **≥90% inclusion of predefined critical aspects**
- **Flesch Reading Ease Score ≥70**
- **Logical structural coherence across key sections**

<REFERENCES>

[1] **Rao, A., Aithal, S., & Singh, S.** (*2011–2023*). *Single-Document Abstractive Text Summarization: A Systematic Literature Review*. Manipal Institute of Technology, Manipal, India.

[2] **Faber, W., Bootsma, R. E., Huibers, T., van Dulmen, S., & Brinkkemper, S.** (2024). *Comparative Experimentation of Accuracy Metrics in Automated Medical Reporting: The Case of Otitis Consultations*. Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands.

[3] **Mathews, D., Varghese, J. P., & Samuel, L. C.** (2024). *Classifying AI-generated Summaries and Human Summaries Based on Statistical Features*. Department of Data Science, Christ (Deemed to be University), Lavasa, India.

[4] **Merine, R., & Purkayastha, S.** (2022). *Risks and Benefits of AI-generated Text Summarization for Expert-Level Content in Graduate Health Informatics*. Dept. of BioHealth Informatics, School of Informatics & Computing, IUPUI, Indianapolis, IN, United States.

[5] N. Rane and S. Govilkar, “Recent trends in deep learning based abstractive text summarization,” Int. J. Recent Technol. Eng., vol. 8, no. 3, 2019.

[6] H. S. Moiyadi, H. Desai, D. Pawar, G. Agrawal, and N. M. Patil, “Nlp based text summarization using semantic analysis,” International Journal of Advanced Engineering, Management and Science, vol. 2, no. 10, p. 239678, 2016.

[7] P. Verma and A. Verma, “A review on text summarization techniques,” Journal of Scientific Research, vol. 64, no. 1, pp. 251–257, 2020.

[8] I. Mani and M. T. Maybury, “Advances in automatic text summarization mit press,” 1999

[9] M. ter Hoeve, J. Kiseleva, and M. de Rijke, “What makes a good summary? investigating the focus of automatic summarization in an educational context,” arXiv preprint arXiv:2012.07619, 2020.

[10] H. Saggion, “Automatic summarization: an overview,” Revue franc¸aise de linguistique appliquee´ , vol. 13, no. 1, pp. 63–81, 2008.

[11] S. H. B. Sri and S. R. Dutta, “A survey on automatic text summarization techniques,” in Journal of Physics: Conference Series, vol. 2040, no. 1. IOP Publishing, 2021, p. 012044.

[12] F. Kiyoumarsi, “Evaluation of automatic text summarizations based on human summaries,” Procedia-Social and Behavioral Sciences, vol. 192, pp. 83–91, 2015.

[13] N. Lehto and M. Sjodin, “Automatic text summarization of swedish ¨ news articles,” 2019.

[14] N. Kobis and L. D. Mossink, “Artificial intelligence versus maya ¨ angelou: Experimental evidence that people cannot differentiate aigenerated from human-written poetry,” Computers in human behavior, vol. 114, p. 106553, 2021.

[15] D. Scott, C. Hallett, and R. Fettiplace, “Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories,” Patient education and counseling, vol. 92, no. 2, pp. 153–159, 2013.

[16] S. Yamamoto, R. Suzuki, T. Fukusato, H. Kataoka, and S. Morishima, “A case study on user evaluation of scientific publication summarization by japanese students,” Applied Sciences, vol. 11, no. 14, p. 6287, 2021.

[17] “Benefits of automatic text summarization,” Oct 2020. [Online]. Available: https://neconnected.co.uk/ benefits-of-automatic-text-summarization/

[18] O. Klymenko, D. Braun, and F. Matthes, “Automatic text summarization: A state-of-the-art review.” in ICEIS (1), 2020, pp. 648–655.