Final Report

20234583 Bekzat Tilekbay
20233716 Hyoungwook Jin
20225492 Kihoon Son

Overview

Data privacy has emerged as a critical issue in the digital age, where the ubiquity of internet technology intersects with the vast expanse of personal and sensitive information available online. In recent years, the significance of safeguarding this data has been propelled to the forefront of public consciousness, primarily driven by the aggressive tactics employed in data crawling and collection. These methods, often employed by companies and organizations, pose significant risks to individuals' privacy and personal security.

This situation is further complicated by the general public's lack of engagement with the terms and conditions (T&C) agreements that govern the use of online platforms and services. Despite their critical importance, these documents are often overlooked or skimmed through without due diligence. Lengthy and laden with legal jargon, these agreements contain key information about how personal data is collected, used, and shared. The failure to thoroughly understand these terms leaves individuals vulnerable to privacy breaches and exploitation of their personal information.

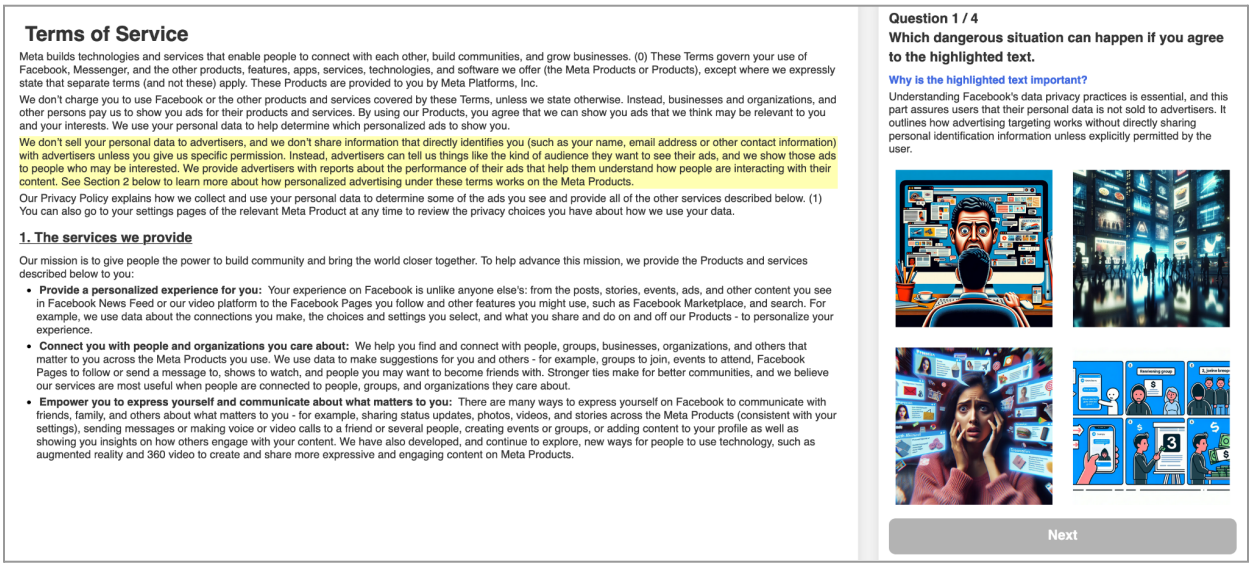

Figure 1. Our interface to help users understand terms and conditions agreements. On the left, terms are displayed with highlighted parts. On the right, our interface asks users to solve a multiple-choice problem to check their understanding.

In this project, we sought to explore whether generative models can play a pivotal role in enhancing the comprehension of Terms and Conditions (T&C) agreements, thereby enabling individuals to more accurately anticipate the potential risks associated with consenting to these documents. To this end, we developed an interface tailored for T&C agreements. This interface utilizes advanced generative models, specifically GPT-4and DALL·E 3, to achieve three key objectives: (1) to underscore the crucial segments of the T&C documents; (2)to generate a variety of multimodal questions aimed at verifying the reader's understanding of these critical sections; and (3) to depict possible dangerous outcomes through imagery. This approach attempts to make T&C agreements more accessible and understandable, potentially mitigating the risks associated with uninformed consent to online data practices.

Implementation

Our interface wasconstructed using Next.js, a comprehensive web framework that facilitates both client and server-side web development. This framework, built on the foundations of React.js and Node.js, offers a robust and flexible environment for web application development. To effectively identify key sections within T&C agreements and generate pertinent questions, we employed a sophisticated pipeline integrating GPT-4 and DALL·E 3, alongside LangChain package for efficient document retrieval and the effective chaining of prompts. This combination of technologies ensures a high degree of precision and relevance in processing and presenting T&C content, thereby enhancing the overall user experience and understanding of these complex documents.

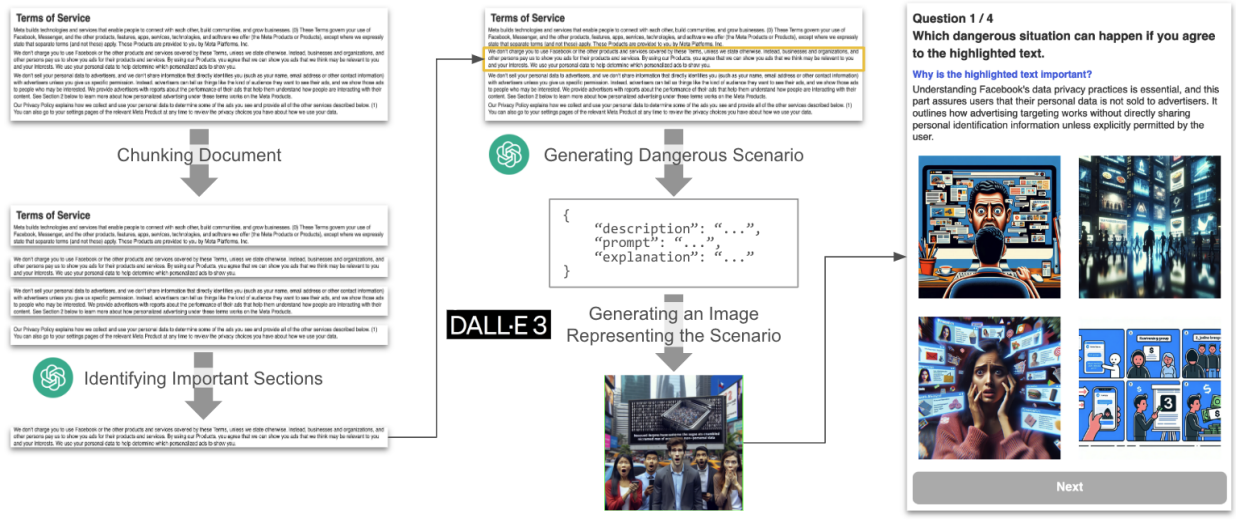

Figure 2. The overview of our pipeline to extract important sections in a terms and conditions document and generate critical thinking questions.

For the scope of this academic project, our implementation currently supports only a limited selection of pre-loaded T&C documents, constrained by the project's timeframe and scope. The front-end interface of our system is not yet configured to accept user-submitted T&C documents for analysis. However, it's important to note that the underlying pipeline of our system is designed to be document-agnostic, offering potential for future expansion and adaptation. A

promising direction for further development could be transforming our system into a Chrome extension. This extension could potentially allow users to receive tailored questions about any T&C document they encounter on various websites. To facilitate ongoing research and development in this area, we have made our codebase and pipeline implementation available to the public on GitHub. The project can be accessed at the following link for those interested in building upon our work: https://github.com/jhw123/cs-projects/tree/ethics.

Evaluation

Goal

This study aims to explore how does the types of quizzes affect the terms and policy document reading experience. We plan to evaluate our approach against plain reading of terms & conditions and compare how well did participants understand the document (RQ1) and the perceived confidence of users in their decision (RQ2).

Study design

We recruited two participants for our study. Our study is based on a within-subject format, and we defined the baseline as the system as a system that support T&C reading with the summary quiz (Figure 3). The treatment system is support the reading T&C with the image quiz. The study started with a brief introduction (3 minutes), and we gave the participants two T&C reading tasks (20 minutes). One T&C is came from Google and the another one is came from Facebook. The task conditions were randomly assigned, and participants conducted a survey after each condition of the task was done. After the tasks, we conducted a follow-up interview session (5 minutes). To answer the research questions above, we defined measures as follows:

Measure (Survey)

- Decision Confidence
- Quiz Difficulty
- Quiz Helpfulness
- Quality of the Quiz

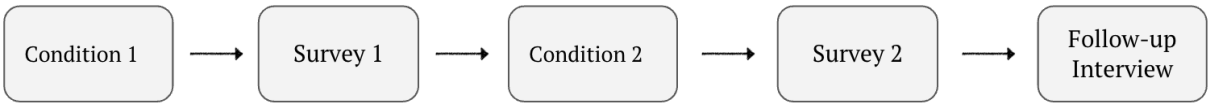

Figure 3. User study procedure

Results

The results of the user study can be summarized as two results. In the Baseline (text quiz) condition, users are more confident in their final decision, but the perceived difficulty is lower in the treatment condition (image quiz). As you can see in the table below, participants expressed more confidence when they answered the text quiz with more difficulty.

Self-reported	reading with a summary quiz	reading with image quiz
Confidence	**5.50 (SD=0.71)**	4.00 (SD=2.83)
Difficulty	**4.50 (SD=3.54)**	3.50 (SD=3.54)

Figure 4. Self-reported confidence and difficulty in the survey.

Second, the text quiz is more helpful with reading, but the perceived quality of the quiz is higher in the treatment condition. Participants reported that the text condition is more helpful, but the quality of the quiz, which includes images, is better.

Self-reported	reading with a summary quiz	reading with image quiz
Helpfulness	**4.50 (SD=0.71)**	3.00 (SD=2.83)
Quality	2.00 (SD=1.41)	**4.00 (SD=2.83)**

Figure 5. Self-reported helpfulness and perceived quiz quality in the survey.

From these results, we can summarize three main findings from our study: First, both quizzes (summaries and images) nudge users to critically assess the T&C and decide whether to accept them or not. Second, when the dangerous consequences are illustrated well within images, users tend to think more critically about the T&C. Lastly, although the textual information is challenging to understand compared to the images, completing the quiz can make them reconsider the document when they make the final decision.

Discussion

Enabling Critical reading of T&C

The study results showed that our interface with two types of multiple-choice questions was effective in enabling critical reading of T&C documents. The contrast between the two types of MCQs indicates that it is important to consider the format of the questions to balance the trade-off between difficulty and helpfulness. Additionally, it is crucial to reduce the cognitive load of the users when going through MCQs, as users should not memorize the document, but only learn about what they are agreeing with and anticipate what can happen. In this sense, image MCQs seem to be the best approach as they allow users to more directly learn possible

dangerous consequences without delving into intricate differences between generated summaries.

Utility of Generative AI

We employed Generative AI models to (1) extract important parts of the document, (2) generate correct and incorrect summaries, and (3) generate scenarios of dangerous consequences of agreeing to the document and respective images for each scenario. We found that GPT-4 and DALL*E 3 work reasonably well in our scenario and the generated MCQs were of good quality. And as Generative AI models get better, the approach should become more practically plausible.

Some stages of our pipeline have a more significant effect on the final quality of the MCQs. Missing important parts of a T&C document can have a significant impact on what the user can learn. It might be important to involve a qualified person in the process to ensure that all the parts are captured well. However, generating slightly wrong summaries or images (e.g., hallucinations) can have a reverse effect on critical reading and we found that users pay more attention in those cases. It might be due to users sensing something wrong with the question or the response options and re-reading the important part of the T&C document several times. Despite these positive effects, it is still important to ensure good quality of MCQs as well as coverage of all the important parts of the document. Incorporating reporting functionalities and *the "none of the options apply"* option might be reasonable approaches to ensure quality.

There is also potential to construct datasets from MCQs similar to approaches like CAPTCHA. The correct summaries & images could be used to evaluate the reasoning capabilities of new Generative AI models.

Ethical aspects

We want to emphasize that this type of quizzing has to be handled by third parties, not the owner company of the T&C document. Due to conflicts of interest, the companies might game the system and conceal important parts of the document or provide wrong options to mislead the users. For this cause, it might be important to organize communities around T&C documents that will develop the quizzing system and ensure its effectiveness.

Limitations & Future Work

We experimented with two types of MCQs: best summary, and best image. We chose these options as they are appropriate for T&C document reading scenarios and can be automatically generated with Generative AI models. As in most MCQs, we also opted for 1 correct and 3 incorrect options strategy without any penalties for wrong answers. However, we leave to future work to investigate more variety within these variables and different types of tests such as free-form responses.

Although our pipeline is useful for T&C reading scenarios there are potentially many ways for further enhancements and optimization. Currently, the prompts that we used to generate the summaries and images are rather primitive and we did not experiment with different documents. Thus, prompt engineering can be done on specific steps of the generation to experiment with the generation outcomes and overall quality. For practical cases,

pre-generating the MCQs for multiple documents for consistency and verifying with real T&C experts can significantly optimize the pipeline and the system.

In terms of evaluation, we did not include a technical evaluation of the generation pipeline due to the manual effort and expertise required to go through generation outcomes. However, it is important to perform technical evaluations on various representative documents to ensure the safety and efficacy of our method. Moreover, our user evaluation was done with only 2 participants, so expanding the studies with a larger and more diverse sample size for robust conclusions is necessary.