TermSight: Making Service Contracts Approachable

Ziheng Huang
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
zihengh2@illinois.edu

Tal August
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
taugust@illinois.edu

Hari Sundaram
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA
hs1@illinois.edu

Abstract

Legal contracts govern much of our society, but their specialized language is difficult for non-experts to read. While AI has enabled simplification of complex language, legal contracts pose unique challenges because of their connection to readers' values, ambiguity, and legally binding nature. Based on a formative study (N=20) using Terms of Service (ToS) as example contracts to study challenges in contract reading, we developed TermSight, an intelligent reading interface to probe the opportunities and challenges of designing augmentations for legal text. TermSight guides readers to relevant clauses with color-coded plain-language snippets of information and contextualizes ambiguous language with definitions and hypothetical scenarios. Importantly, TermSight's features always foreground the original, legally-binding contract text (e.g., linking to associated clauses). Our within-subjects study (N=20) demonstrated the opportunities of TermSight in making ToS significantly easier to read and navigate while revealing the challenges of augmenting service contracts such as ToS.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Law**.

Keywords

Legal Contracts, Augmented Reading, Large Language Models

1 Introduction

Legal contracts govern much of our society. We sign contracts for where we live, who we work with, and, increasingly, for our digital interactions online. In order to encompass many possible situations and actions, contracts contain highly specialized language, such as:

> *"You grant us a worldwide, royalty-free, perpetual, irrevocable, non-exclusive, transferable, and sublicensable license to use, copy, modify, adapt, prepare derivative works of, distribute, store, perform, and display Your Content."* — Reddit Terms of Service [3]

On the one hand, this specialized language allows contracts to articulate the entitlements, obligations, and prohibitions [89] of the signing parties. On the other hand, this specialized language is exceptionally difficult for many people without a legal background to read [6, 33, 74].

Recent AI systems, driven by language models (LMs), have expanded our ability to transform and augment text documents for different readers. Past reading systems have simplified scientific papers [92], medical research [8], and news articles [21] for general audience readers. These transformations have also begun to be applied to legal language, by, for example, generating summaries based on predefined categories of rights and responsibilities in contracts [82, 89] or through explaining legal concepts [50].

However, legal text, and specifically contracts, pose unique challenges for simplification that current systems fail to address. First, contracts are inherently value-dependent: they are intended to encode and protect the values of the signing parties [29, 39, 89], which may lead to varying information needs among individuals. Second, legal language is not just complex, but can be ambiguous (e.g., *"We may share some of your personal information..."* [14]). While this language provides the linguistic flexibility to manage unforeseen circumstances [109], it also derails general audience readers [14, 48, 61, 99, 107]. Finally, the literal language of a contract is important: it is the only language that is legally binding [23]. Consequently, any text transformations should support rather than replace reading the original text.

In this paper, we explore augmentations that account for the unique challenges of simplifying legal contracts. We focus on Terms of Service (ToS), the ubiquitous but rarely-read contracts that govern the use of digital services [5, 35, 80]. Despite consumers' concern for the information presented in ToS [73, 79], these contracts remain difficult to engage with. While prior work often focuses on a single policy within the ToS (e.g., a privacy policy [77]), we explore challenges readers face when trying to engage with the full ToS contract. In a formative study (N=20) where participants read ToS, we found that these contracts represented the extremes of the difficulties that legal contracts pose more generally (i.e., value-dependent, ambiguous, and legally binding), validating this initial focus. The parts of the ToS that were important to participants depended heavily on their envisioned usage and personal values, and this value-dependent information was lost among the many nested policies. ToS also contained complex and ambiguous terms that confused readers. Finally, existing features to support contract reading (e.g., policy names and section headers) often did not accurately represent the actual text and were not legally binding.[1]

In response, we designed a novel intelligent reading interface for ToS contracts, TermSight. In contrast to prior work that provides overview summaries detached from the original text [1, 2, 82],

[1]ToS often include clauses such as *"Headings are used in these Terms for reference only and will not be considered when interpreting them"* [3].

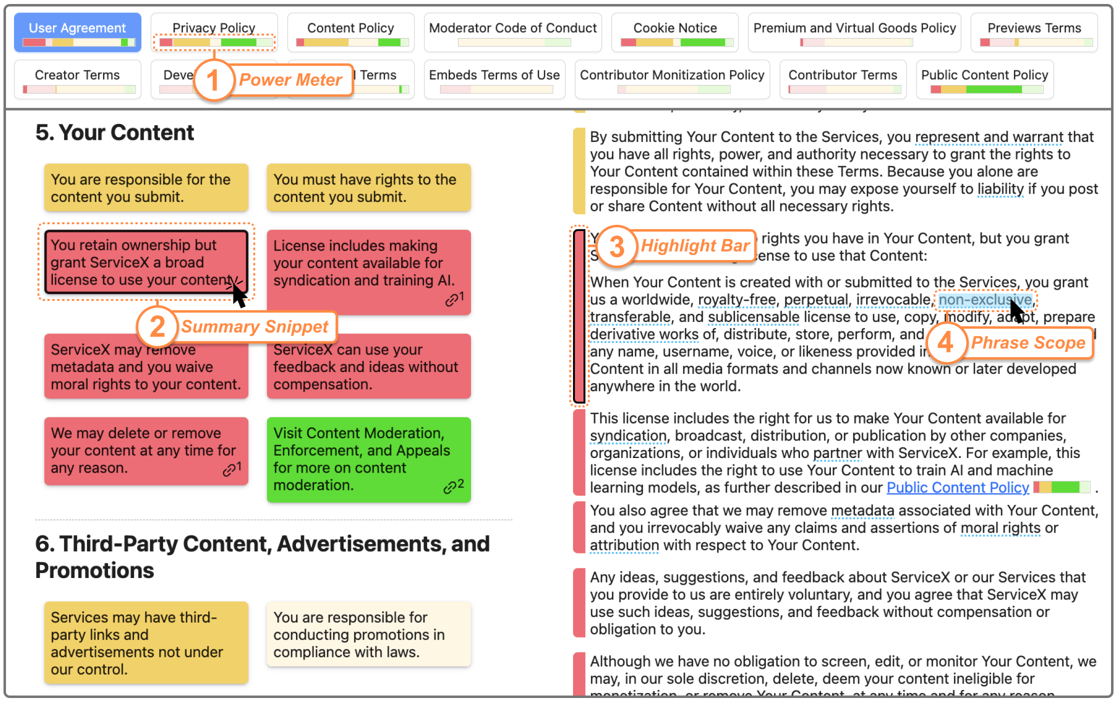


Figure 1: TermSight provides multi-level support for reading Terms of Service (ToS). At the contract level, TermSight visualizes the relevance and power balance of content in each policy (1). At the document level, TermSight chunks, summarizes, and categorizes content into one-sentence plain-language summaries (Summary Snippets) highlighted with colors that reflect power and relevance (2). Readers can click the Summary Snippet (2) or Highlight Bar (3) to navigate between the two. At the phrase level, TermSight offers phrase definitions and hypothetical scenarios for unfamiliar and ambiguous phrases (4).

we explore augmentations that support contract reading while acknowledging the unique characteristics of legal text. TermSight provides visual overviews of policies within a ToS (Power Meter, Figure 1.1), guides readers to relevant clauses with color-coded plain-language snippets of information (Summary Snippets, Figure 1.2), and contextualizes unfamiliar or ambiguous language with definitions and scenarios of potential implications (Phrase Scope, Figure 1.4). Importantly, TermSight's features always foreground and link to the original legal text.

To explore the opportunities and challenges of designing contract reading interfaces, we conducted a counterbalanced within-subjects study (N=20) using TermSight and a baseline HTML reader. Compared to prior focus on simplified reading tasks (e.g., reading a single clause [62, 88] or policy overview [57, 97]), participants engaged with real ToS contracts with 10+ policies. In the study, participants found reading ToS to be significantly easier with TermSight while also being more willing to read the original ToS. Participants reported that TermSight provided multi-level guidance and support while letting them drill down into clauses within the ToS important to them. We also observed that participants often used TermSight's generated text (e.g., Summary Snippets) as an entrance to, or a verification of, the original ToS text. At the same time, our findings revealed the broader limitations of technological solutions like TermSight on facilitating a deeper understanding of ToS due to the overwhelming amount of potentially relevant information. We end by discussing the implications of our findings for inspiring future contract interfaces and regulatory innovation. In summary, this paper makes the following contributions:

(1) We characterize the challenges of contract reading, specifically ToS, including determining which policies to read, resolving ambiguous terms, and navigating misleading organization within the text.
(2) We develop a novel intelligent reading interface, TermSight, to explore the design opportunities and challenges for helping readers engage with ToS while focusing on the original, legally binding language.
(3) We collect empirical evidence from our user study demonstrating the values of TermSight's features in finding relevant information, navigating ambiguous or complex language, and associating generated text with original text, while revealing the challenges of improving understanding of ToS.

2 Background and Related Work

2.1 Service Contracts

A Terms of Service (ToS) is a type of standard form contract consisting of a body of policies and conditions outlined by a service provider, dictating the rules and expectations for both the service provider and the consumer. While prior work in HCI mainly focused on single policies within the ToS (e.g., a privacy policy [77]), these contracts encompass a wide range of provisions such as copyright, privacy, returns, acceptable use, and service-specific terms [6]. Important terms are often buried within this large body of documents [94]. ToS can vary in content [29, 64] and structure [48] across services. The language used to write ToS can also require advanced reading capability [6, 94] and contain unfamiliar [48, 99] or ambiguous terminology [14, 61, 107]. Most consumers do not read modern ToS [9, 80], despite the binding nature of ToS [28] and consumers' self-reports indicating concern for the information presented in ToS [73, 79]. This widespread disengagement highlights what Kar and Radin described as a "paradigm slip" in contract law, where agreements once premised on shared understanding and mutual consent are now reduced to unilateral boilerplate "pseudo-contracts", eroding fundamental principles of contract law and personal autonomy [12, 52, 81]. Our formative study further supplements prior findings by revealing the challenges readers face when navigating the entire ToS and the presence of misleading navigational affordances (§3).

2.2 Augmenting Terms of Service

Prior work has investigated new designs for readers navigating ToS [40, 54, 98]. Kay and Terry [54] proposed Textured Agreements that use typographic manipulation, pull quotes, vignettes, and iconic symbols to improve reader attention and comprehension of privacy policy compared to plain text. Habib et al. [40] investigated how icons and linked text could be designed to better convey privacy choices. Taber et al. [98] proposed crowdsourced sentiment highlighting of sentences in ToS. Design recommendations on styling user agreements have also been proposed [40, 54].

Work has also explored helping consumers gain an overview of specific policies within a ToS, often the privacy policy (PP), without requiring reading the original document. Platform for Privacy Preferences (P3P) was an early attempt to standardize privacy policies by allowing service providers to submit privacy policies in a machine-readable format [86]. Tools such as Privacy Bird build atop the P3P format to provide warnings when a site's PP does not match consumers' privacy preferences [26]. However, P3P was not widely adopted by service providers [25] and is no longer supported [72]. Work has also proposed the use of a single-page summary for end-user license agreements (EULAs) [35, 36], comic-based summaries for privacy policies [97], and Privacy Policy Nutrition Labels (PPNL) that label privacy policies in a table format [55, 87].

Other work has focused on providing overview summaries automatically or via crowdsourcing. ToS;DR uses volunteers to label and summarize terms in the privacy policy and the home page of ToS [2]. Works like CLAUDETTE automatically detects a predefined set of potentially unfair clauses from ToS [17, 18, 37, 63]. PrivacyCheck [77, 78, 108], PrivacyGuide [100], and Polisis [41] extract terms relevant to a list of privacy-related questions and criteria and generate a grading for each. Past work has relied on either service providers or volunteers to make stylistic changes to contracts or automated tools to provide a summary external to the contract. In this paper, we explore new ways of combining these two threads to enable readers to navigate and read the original ToS made possible by the powerful text transformation capabilities of language models.

2.3 Augmented Reading Interfaces

Prior work has explored augmenting other types of documents, such as research papers [65] and news articles [21], with generative features. Work has provided faceted highlighting in research papers to support skimming [31] and context-sensitive definitions of terms and symbols [44] to support reading. Also in research papers, Qlarify [30] enabled readers to expand paper abstracts with generative explanations. Systems have also provided automated question answering for research papers [110] and medical papers [8]. Shin et al. [92] transformed research papers into summaries of design implications. Newman et al. [75] developed tools that synthesized multiple research papers to help researchers conduct literature reviews. In the business domain, Marco allows readers to search and ask questions across collections of business documents [32]. In the news domain, Chen et al. [21] introduced Marvista, a reading interface that identifies the most summative portions of news articles.

In contrast to other documents explored previously, contracts—and specifically ToS—pose unique challenges to readers. Contracts directly connect with a reader's personal values: one reader might prefer stronger privacy safeguards while another might care about licenses [29, 39, 89]. Contract language is also meant to be interpreted; this means there are often ambiguous terms or incomplete information (e.g., the definition of 'information' when discussing personal information collection), with the assumption that the reader has the requisite knowledge to fill in missing information [14]. Finally, contract language itself is the only legally binding text [23], meaning any text transformations must maintain and display a tight connection with the original text. We take inspiration from past augmented reading systems to investigate what features can support readers in the unique context of contract reading.

3 Formative Study

We first establish the challenges general-audience readers face when reading a legal contract. We focus on Terms of Service (ToS) because ToS are legally binding contracts that are signed routinely in digital interactions. While ToS are rarely read completely [5, 35, 80], past legal work suggests that these contracts represent many of the broader issues faced in legal contracts today [52, 80] and therefore present an ideal setting to explore possible reading features. Our formative study was guided by the following research questions:

> **RQ1**: What information do readers want from a ToS?
> **RQ2**: What are the challenges readers encounter when trying to get this information from ToS?

3.1 Methodology

3.1.1 Study Procedure. The entire study took 40 minutes, and participants were compensated 10$. The study was approved by the Institutional Review Board (IRB) office at our organization. Participants first reported their past experiences interacting with ToS. Next, participants were asked to imagine as if they were a first-time user registering for a randomly assigned service and were provided 15 minutes to go through the ToS while speaking aloud any challenges. Participants were free to navigate to different policies within the ToS. After reading the ToS, we conducted semi-structured interviews about participants' experiences and challenges with ToS reading, along with opportunities for computational support. The interview questions are listed in Appendix D.

3.1.2 Analysis. We conducted a reflexive thematic analysis on the interview transcripts to identify common challenges and themes. We followed the six phases of reflexive thematic analysis suggested by Braun and Clarke [19]. The lead researcher thoroughly explored the data, noted interesting features, systematically coded the data, and iteratively compared the codes to generate the initial themes. Through weekly meetings over the course of two months, the lead researcher discussed with other members of the research team to make sure the themes fit the data and further developed the themes. This included iterating on the descriptions of the themes and merging related themes.

3.1.3 Materials. Participants were randomly assigned one of the 10 services' ToS[2]. The 10 services were selected from the top 20 visited sites on semrush.com (2024/05), where we sampled the top 4 social media services, top 2 E-commerce services, top 2 video platforms, and top 2 internet services. We oversampled social media services because they were the most represented service category in the top 10 visited services.

3.1.4 Participants. We recruited 20 participants through Prolific (Male: 6, Female: 14). All participants were over 18, fluent in English, and located in the US. Participants' age ranged from 20 to 67 (μ = 35.75, σ = 12.24). 3 participants had never read or skimmed a ToS, and 17 participants had read or skimmed at least 1 ToS before. We did not require participants to have prior experience in reading the ToS. All the participants had not read or did not remember reading the ToS they were assigned.

3.2 RQ1: What information do readers want from a ToS?

3.2.1 ***Participants have diverse, value-dependent information needs when reading ToS.*** All 20 participants described the need to know what control and rights the service had over the user, what the service was allowed to do, and how the user could be negatively impacted. While all participants shared a desire to know their general rights, different participants also noted more specific information needs, such as data collection and usage (11), intellectual property rights (8), purchasing and returns (7), account deletion by the service (5), arbitration and liabilities (3), content moderation (2), and exposure to offensive content or misinformation (1). 10 participants explicitly mentioned how their information needs from the ToS and priorities depended on service usage and personal values. For example, P1 described that *"I don't think any information is irrelevant. I just think some parts are more of a priority of knowing. If I happen to become a person who wants to create graphic designs, I would definitely utilize the Copyright portion of the ToS to ensure that I'm being compliant with [the platform's] expectations."* On the other hand, P3, with a background in software, raised privacy concerns when using services for professional communication: *"For Microsoft Teams and Gmail, I was using them for a collaborative project with a small company. I wanted to ensure that whatever data or information was shared would remain proprietary to us."*

[2]Social Media: Reddit, Facebook, Instagram, Twitter; E-commerce: Amazon, eBay; Video: Youtube, Netflix; General: Google, Yahoo

3.3 RQ2: What are the challenges readers encounter?

Our findings revealed that participants encountered challenges at the contract (§3.3.1), document (§3.3.2), and phrase (§3.3.4) level. These challenges reflected the complexity and ambiguity of legal language. In addition, participants highlighted how existing navigational and reading affordances were often misleading and failed to support meaningful sensemaking of the actual ToS text (§3.3.3).

3.3.1 ***Contract Level: Participants struggled to navigate nested policies and decide which policy to read.*** Eleven participants mentioned that it's unclear to them what sub-policies are included when they are navigating the ToS because sub-policies are often hyperlinked throughout the documents. There often lacks a centralized list of what policies are included as part of ToS. Seven participants described the challenge of navigating across these nested policies as getting *'lost in a whole tree of information'* (P18). All 20 participants described relying on policy names to help decide whether to explore a policy. Yet, 17 participants noted that the names of the policies alone did not provide a clear mental model of what was in the sub-policies and whether there was important information they should know about. For example, P14 mentioned that: *"When it's hyperlinked like 'our rules and policies' and 'privacy policy', it's difficult to determine if I need to know something from the link."* 8 participants explicitly expressed their reluctance to click into the linked polices due to the lack of awareness of relevant information and the challenge of finding relevant information from a hyperlinked document.

3.3.2 ***Document Level: Participants lacked guidance and struggled to surface relevant information from extended, visually dense, and obfuscating text within a policy.*** Most (17/20) participants primarily skimmed the ToS. Six participants noted that they were not sure what they should look for in the ToS and might have unintentionally missed important information. P12 further commented how it's not always obvious if the information was relevant at first sight: *"A lot of it seemed irrelevant until I started reading in a little bit more."* Often, this inability to find information was connected to the overwhelming length of the document (18/20). Interestingly, many (13/20) participants also noted how text positioning seemed intentionally obfuscating: important information was positioned at the bottom of an extended policy. For example, P10 noticed how information about returns and cancellations of an E-commerce platform was positioned under the 13th section with the title 'Additional Terms'. Similarly, three participants described how the use of friendly language tended to obfuscate information and made skimming more challenging. For example, P15 noticed language such as *'We only use data to make [the service] a better place'* when reading a privacy policy. Yet, she later found out how user information can be used for targeted ads.

3.3.3 ***Document Level: Participants found existing affordances to be ineffective and misleading.*** Some of the ToS participants read contained navigational and reading affordances, such as interactive table of contents with section headers and overview summaries. However, 12 participants noted that, similar to policy names, section headings within policies were vague and not descriptive of the actual content. For example, when skimming the ToS, P10 skipped the 'Content' section but later realized that the section was about intellectual property rights over user-generated content: *"It just says 'content' which is pretty vague when they're talking about getting exclusive intellectual rights. I feel like it's misleading. So people would skip right past it. I did the first time."* In addition to ineffective navigational affordances such as vague policy names (17/20) and section headers (12/20), three participants accessed a sub-policy with a summary at the top. All three participants (P4, P14, P15) described using the summaries to gain an overview at the start. Yet, after reading the policy, participants noticed that the summary was missing key details: *"The summary is kind of misleading. They gave you a more palatable version of [ToS]. If you scroll down, it will say we can terminate your account for no reason. They don't say that in the summary"* (P4).

3.3.4 *Phrase Level: Participants struggled to contextualize unfamiliar or ambiguous terminology.*

When reading individual sections in a policy, 12 participants pointed out unfamiliar legal terminology (e.g., *Arbitration*, *Indemnity*, or *class lawsuit*). These participants noted the difficulty to interpret the meaning of legal jargon in the context of their situation: *"I can recognize these are legal terms, but I don't necessarily know what that means for me"* (P10). In a similar vein, eight participants noticed ambiguous terminology that made it challenging for them to understand the implications of signing (e.g., 'we share your data with 3rd parties', and 'retain certain information about you'). P15 mentioned that *"I really think the big difficulty is the vagueness of the language and the constancy of exceptions that are vague, which gives them a lot more leeway"*.

3.4 Design Guidelines

We propose four design guidelines based on our observations that readers faced interrelated barriers at all levels of ToS reading:

> **DG1**: Contract level: Help readers form mental models of each policy of a ToS to determine important documents to interrogate.
> **DG2**: Document level: Help readers identify potentially relevant information within a policy.
> **DG3**: Phrase level: Help readers read and interpret original text with technical language, jargon, and ambiguous phrases.
> **DG4**: All summaries should be tightly coupled with the text used to generate them. When reading any generated overviews, readers should be able to retrieve the original document text in at most one click.

4 The TermSight System

We reify these design guidelines in TermSight (Figure 1), an augmented reading interface for exploring opportunities to make ToS contracts more approachable with features at every granularity:

> **Power Meter:** A *contract-level* visualization of the relevance of information within a document and the distribution of power between the user and the platform that information represents.
> **Summary Snippets:** One-sentence plain language summaries of information chunks at the *document-level*. Snippets are rendered with color and saturation that represent power and relevance to the user, similar to the Power Meter.
> **Phrase Scope:** In-situ *phrase-level* definitions and scenarios via a tooltip that contextualizes the meaning of the phrase while allowing users to ask clarification questions.

We adopted an iterative design approach in developing TermSight. Eight participants evaluated an early prototype of TermSight, and their feedback informed the final design. Additional details about the iterative design study, results, and design changes can be found in Appendix A.

4.1 Design Abstractions

The design of TermSight is grounded in two design abstractions: 1) *Information Snippets*, and 2) *Power* and *Relevance* classification.

4.1.1 *Information Snippets.*

Our formative study revealed that the information readers cared about in ToS can be buried within visually dense text. To enable sense-making of finer-grained pieces of information as opposed to an entire section, we use **Information Snippets** as the unit of interaction in TermSight. An Information Snippet is a continuous span of original text that shares the same topic and can be summarized in a single sentence of up to 12 words (detailed in Section 5.2). These Information Snippets form the basis of the Summary Snippets (§4.3.1). Information Snippets are non-overlapping and together reconstruct the original document's content. Figure 5 illustrates the process of how TermSight divides a paragraph into five Information Snippets, each paired with its Summary Snippet. The idea of designing interactions and sense-making around finer-grained Information Snippets was inspired by prior works that modularize and 'objectify' tools [10, 15], attributes [106], and AI agent's memory [47] to enable interaction specificity and direct manipulation.

4.1.2 *Power and Relevance Classification.*

Our formative study revealed participants' information needs regarding the rights and control held by the service and the user. These needs may also vary depending on the intended service usage. To provide information scent [85] for snippets of information that a reader may care about, we reify participants' information needs by defining **Power** and **Relevance** as two qualities of an information—and thus a summary—snippet. Power is the degree to which a snippet's text grants control to the Service Provider or the User (Categories: Service, Neutral, or User). Relevance is whether the snippet's text is directly relevant to the user's intended usage of the service or their values (Categories: High, Low). For example, a snippet related to selling would have low relevance for a user more focused on buying. As shown in Figure 2, we visually encode Power and Relevance in TermSight with a combination of hue (Red, Yellow, Green) and saturation (High, Low). We constrained relevance to High and Low because our iterative design study revealed that more color–saturation pairs (i.e., 3 hues for Power x 3 saturations for Relevance) were difficult to interpret (Appendix A).

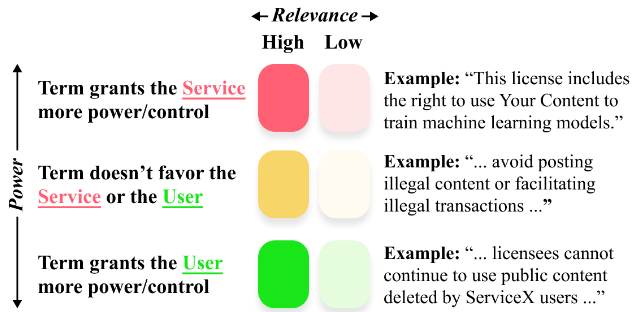


Figure 2: Design conceptualization of Power and Relevance. Power refers to the degree to which a snippet grants control to the service provider or the user. Relevance refers to whether or not the snippet is relevant to the user's persona.

4.2 Power Meter: Visualization of the Distribution of Power and Relevance

In TermSight, the policies that are part of the ToS are placed in the top navigation panel. To help users form mental models of each policy and decide which policies to read (DG1), each policy in the top navigation panel or hyperlinked within the document is accompanied by a **Power Meter**: a horizontal bar visualization representing the distribution of Information Snippets in a document, with colors denoting power and relevance (Figure 3). Our iterative design study further revealed the need to contextualize Power Meter with concrete previews (Appendix A). Consequently, hovering over a Power Meter reveals a policy preview containing a list of Summary Snippets (Figure 3b). Each Summary Snippet in the popup includes a clickable icon (Figure 3c) that allows users to view the referenced Information Snippet (DG4).

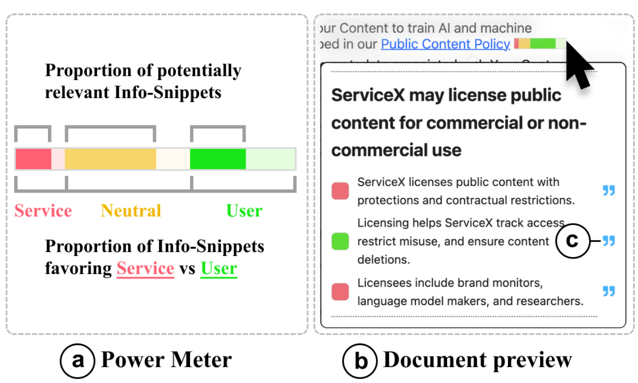


Figure 3: Power Meter visualizes the distribution of power and relevance of Information Snippets within a policy (a). On hover, a preview of Summary Snippets is shown (b) with options to view the referenced Information Snippets (c).

4.3 Summary Snippets: Plain language Summaries of Information Snippets

To help users surface relevant information within a document (DG2) and interrogate the original text (DG4), TermSight provides Summary Snippets accompanied by Highlight Bars for seamless navigation between the summary and the original text (Figure 1).

4.3.1 Summary Snippets. In contrast with document-level [56, 101] or section-level summaries [8, 67], TermSight explores supporting interaction and sensemaking for finer-grained snippets of information within dense sections. TermSight features **Summary Snippets**: one-sentence plain language summaries of Information Snippets (§4.1.1). Summary Snippets are designed to be short (detailed in Section 5.2), which have been shown to be more effective for navigational tasks compared to longer snippets [27, 96]. Each Summary Snippet is rendered with color based on the Power and Relevance of the corresponding Information Snippet to provide information scent [85] for deciding which snippets to pay attention to (DG2). To enable users to interrogate original text and verify AI-generated summaries (DG4), clicking on a Summary Snippet jumps a user to the corresponding Information Snippet in the document.

4.3.2 Highlight Bar. To help users surface and identify Information Snippets while reading or skimming the original text (DG2), TermSight introduces **Highlight Bars**. Highlight bars are positioned to the left of each Information Snippet in the original text, visually segmenting dense sections into Information Snippets. For paragraphs containing multiple Information Snippets, additional line breaks are added after each snippet for visual separation. Similar to the Summary Snippets, Highlight Bars are color-coded based on the Power and Relevance of the corresponding Information Snippet (§4.1.2). Users can click on a Highlight Bar to jump to the corresponding Summary Snippet.

4.4 Phrase Scope: Phrase Identification, Definition, Scenario, and Clarification

To help users engage with the original text (DG3), **Phrase Scope** first identifies phrases that may be unfamiliar or ambiguous to general audience readers and underlines them in blue (Figure 4). This provides a starting point for users to explore potentially unfamiliar phrases without relying on users to identify them. Users can also highlight any arbitrary span of text to request Phrase Scope. Clicking on the system or user-identified phrase opens Phrase Scope, a tooltip with three components. First, the definition of the phrase in the context of the ToS is offered, similar to prior work in other domains [30]. Beyond definitions, participants in the formative study described the challenge to contextualize abstract phrases in unforeseen scenarios. Inspired by prior work that encouraged scientists to reflect on the unintended consequences of their research [83, 104], Phrase Scope explores the opportunity to prompt reflection on the implications of a phrase by offering hypothetical scenarios personalized to the user persona. Lastly, users have an option to ask clarification questions. More examples of the generated definitions and scenarios can be found in Appendix Table 2.

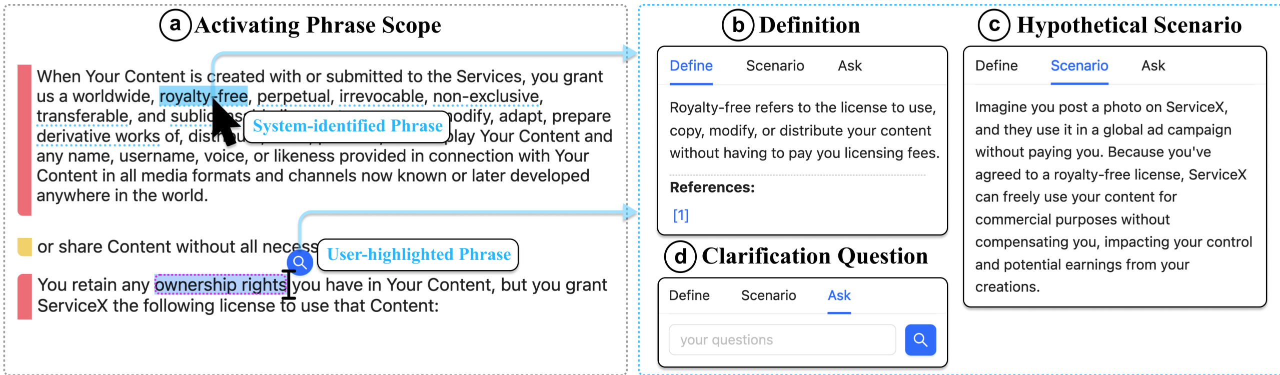


Figure 4: Phrase Scope first identifies jargon or vague phrases in the contract (a). Clicking an identified phrase opens a tooltip with definition (b) and hypothetical scenario (c) while allowing users to ask clarification questions (d).

5 Implementation Details

TermSight is rendered as a web interface implemented with Next.js. Below, we explain the input, output, processing, and overall performance of TermSight. Additional implementation details and prompts are specified in Appendix B.

5.1 System Input and Pre-processing

TermSight assumes a set of HTML or markdown source files representing the policies included in the Terms of Service (ToS). We assumed clean source files because the focus of TermSight is not on developing web scraping technologies but rather on investigating meaningful navigational and reading affordances for ToS contracts. Each document is segmented into text **"chunks"**, composed of one or more paragraphs, which are further vectorized and stored in a vector database (detailed in Appendix B.1). For features that depend on user preferences, such as classifying the relevance of Information Snippets (§5.2) and generating personalized scenarios (§5.3), TermSight uses a text-based persona that includes users' intended usage of the service (e.g., content consumers vs. content creators) and their values or concerns (e.g., privacy, copyright, etc). The user personas used for the study are further explained in §6.1.2 and displayed in Appendix E.2.

5.2 Obtaining Summary Snippets and Classifying Information Snippets

For each chunk of text obtained from the document pre-processing pipeline, we prompted GPT-4o to generate a list of one-sentence summaries (Summary Snippets) each referenced to a span of the input text chunk (Information Snippet) as illustrated in Figure 5. While the length of the referenced text for each summary is unrestricted, each summary is constrained to a maximum of 12 words based on prompt engineering and prior work showing that short summaries can be more effective for navigation (e.g., 10-20 words) [27, 96]. We noticed that setting summaries shorter than 12 words led to excessive fragmentation (i.e., too many Summary Snippets), while longer summaries tried to cover too much information, making them less skimmable. We then prompted GPT-4o to classify each Information Snippet along two dimensions: Power and Relevance (to user persona). Prompts can be found in Appendix B.2 and B.3.

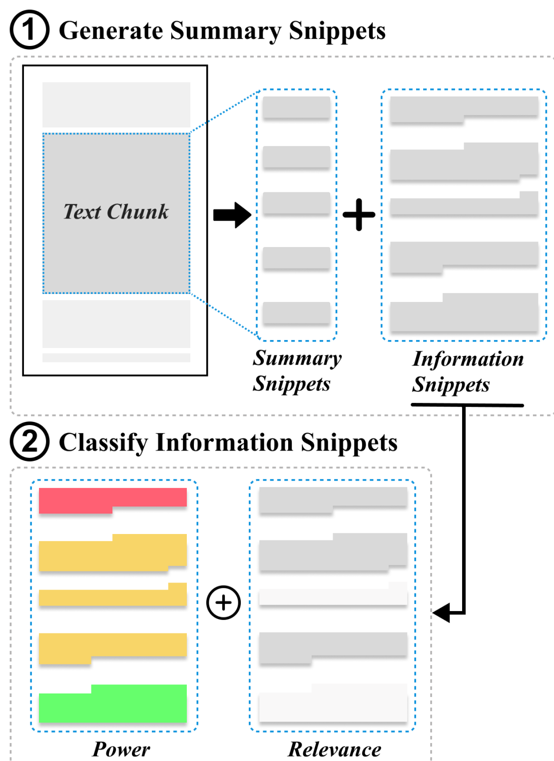


Figure 5: A flowchart of the implementation for (1) obtaining a list of Summary Snippets each referenced to a span of the input text (i.e., Information Snippet) and (2) classifying Power and Relevance for each Information Snippet.

5.3 Generating Phrase Scope

Similar to prior work [30, 38], we prompt an LM to identify potentially unfamiliar and ambiguous phrases within document chunks produced by the pre-processing pipeline. To generate in-context definitions for the identified phrases, we used a retrieval-augmented question answering approach [30]. We vectorize the document chunks (§5.1) and query these chunks to retrieve relevant document context. These chunks are collectively referred to as **"retrieved chunks"**. Then, we prompted GPT-4o to generate an in-context definition with references to the retrieved chunks. When users ask additional questions, the same retrieval-augmented question answering pipeline used for generating definitions is applied, with the only difference being the question asked. Scenarios aim to prompt reflection on the potential implications of a phrase based on users' intended usage of the service and personal value. We employed GPT-4o with zero-shot prompting to generate the scenarios tailored to the user persona. The prompts used for Phrase Scope are detailed in Appendix B.4.

5.4 Evaluation of System Output

Before reporting on our user study, we randomly sampled and conducted a manual evaluation of TermSight's core outputs (e.g., summary snippets, definitions, scenarios). Our goal was not to assess advances in system performance, but to verify that the system can produce meaningful outputs that could support ToS reading. Across 116 sampled Information Snippets, we found 5 imperfect classifications of power or relevance due to the lack of full context in the input Information Snippet. Similarly, we reviewed 113 generated definitions and scenarios each. One scenario was found to be factually incorrect due to a hallucination. The input context and phrase stated that users do not gain ownership rights by downloading content from the service. However, the scenario claimed that users might lose ownership rights over the content they create by uploading it to the service. All the definitions were correct except for 4 overly general definitions for service-specific phrases not explicitly defined anywhere in the ToS. More details of the evaluation are provided in Appendix C.

6 User Study

Similar to prior HCI studies [20, 21, 51], the goal of our user study is not to prove TermSight as the definitive solution through a series of ablation experiments, but to use TermSight to explore the problem and solution space of designing contract reading interfaces. We asked the following research questions:

> **RQ1**: How did participants perceive TermSight and its features?
> **RQ2**: How did TermSight influence ToS comprehension and recall?
> **RQ3**: How did participants read with TermSight and its features?

6.1 Study Design

We conducted a within-subject study where each participant used both TermSight and a baseline interface to read two services' ToS, once with each interface variant. The order of the interface and service type was counterbalanced to reduce ordering effects.

6.1.1 Baseline. Because there is no well-established standard ToS reader, we use the most commonly used interface as the baseline (i.e., an HTML reader). The baseline interface had the same layout as TermSight without the Power Meter, Summary Snippet, Highlight Bar, and Phrase Scope (Appendix Figure 25). In place of the Summary Snippets was a table of contents that can be used to navigate to different sections. Participants' feedback in the user study confirmed that the baseline offered a strong frame of reference (§7.1). While reading interfaces exist for other domains (e.g., [8, 21, 31]), adopting them to contracts would result in novel systems unsuitable as baselines. For example, prior work has designed key questions [8] or faceted highlighting [31] specific to academic papers, which require co-design studies with lawyers to adapt them to contracts. Injecting additional features (e.g., overview summaries [97], chat [110]) into one or both conditions may also introduce confounds unrelated to our research questions. For example, by adding a policy or section-level summary in the baseline, the comparison with TermSight will be confounded by the design of summaries in the baseline (e.g., levels of complexity and detail), a variable that may impact different readers differently [7].

6.1.2 Materials.

Terms of Service. For the user study, we used the ToS of one social media site (Reddit) and one e-commerce site (Poshmark) because social media and e-commerce platforms are among the most visited digital services and are representative of the standard long-form contract used in prior studies [35, 54, 80, 97, 98]. For each service, we collected policies that are linked as part of the ToS. We did not include location-specific policies (e.g., a California Privacy Notice). For Reddit, we also did not include policies for advertisers, publishers, or governmental agencies to keep the number of policies for both services similar. We collected a total of 14 policies for Reddit and 15 for Poshmark. The service names for Reddit and Poshmark were anonymized as ServiceX and ServiceY.

User Persona. Features of TermSight rely on a user persona to determine relevance. For the user study, we designed two personas based on users' information needs identified in the formative study: a content consumer who posts personal content on social media sites and a buyer who rarely posts reviews on e-commerce sites (Appendix E.2). Participants in the user study validated and found their given user persona for both services to highly align with their personal usage of the service and their personal value (Figure 13).

6.1.3 Study Procedure. The entire study took 90 minutes, and participants were compensated 25$. The study was approved by the Institutional Review Board (IRB) at our organization. The study was composed of two reading sessions, 45 minutes each. After obtaining consent and before the reading sessions, we asked participants about their past experiences with ToS. Within each reading session, participants first completed a pre-survey and reported their past experiences interacting with social media or e-commerce platforms. Then, an anonymized description of the service and user persona was given. Similar to prior studies [8], participants were given 10 minutes to read the ToS without having to speak aloud. Reminders were given at the 5 and 1-minute mark. After the reading session, participants completed seven 5-point Likert-style rating questions about their reading experiences, followed by one free-response recall question and six multiple-choice comprehension questions. We also conducted semi-structured interviews with participants about their experiences and reading strategies for 10 minutes. The same procedure was repeated for the second reading session. The study materials are included in Appendix E.

6.2 Participants and Setting

We recruited 22 participants through Prolific. The recruitment required participants to be over 18, fluent in English, located in the US, not have color deficiencies, and not be legal professionals. We did not require participants to have prior experience in reading ToS. Two participants encountered technical difficulties, and their data were removed from the analysis. The median of participants' self-rated familiarity with ToS was 3.5 ($\sigma = 1.5$) out of 5. Most participants were generally familiar with social media and e-commerce platforms and found their given user persona to highly align with their own profile (Figure 13). None of the participants had read or remembered the ToS for Reddit and Poshmark prior to the study. Additional participant demographics can be found in Appendix E.3.

6.3 Measures

6.3.1 Reading Experiences. We collected seven 5-point Likert-style ratings (1="Not at all," 5="Very") of participants' reading experiences (Appendix E.5). The experience measures included ease of reading (adopted from perceived effort measure in NASA TLX) [8, 43], perceived understanding [8], and confidence in obtaining relevant information [8]. We added two questions specific to challenges identified in our formative study: the ease of deciding which policy to read and what text to read within a policy. Finally, we included two questions on participants' willingness to read ToS with the interface and their willingness to spend more time on the ToS with the interface [7].

6.3.2 Comprehension. We wrote 10 multiple-choice questions for each service and selected 6 questions that are relevant for the given user persona (§6.1.2) for each service, similar to prior work on evaluating reading comprehension for ToS and privacy policies [8, 97, 102]. We designed the comprehension questions to have one answer based on the original text, independent of features of TermSight (e.g., Summary Snippets). Consequently, participants need to read the original text to fully answer the questions. The questions were different for the two services to minimize learning effects. Comprehension was measured as the number of questions participants got right. We note that past studies with reading interfaces often fail to show differences in how people answer comprehension questions [8, 44, 57, 88, 97]. Similar to prior work [8], our intention in including comprehension questions was to ensure that TermSight did not detract from overall comprehension. The questions can be found in the supplemental materials.

6.3.3 Recall. For recall, participants were asked to write down information in ToS they found interesting, surprising, or any detail they remembered: *"What did you learn from reading the ToS? Recall one or more interesting things you learned from the ToS."* To analyze these free-form responses, we manually broke the response into single references to a clause in the ToS. Then, we labeled whether each reference was correct based on the ToS. Recall is measured as the number of correct facts in the response [35, 98].

6.3.4 Feature Usage. To understand users' interaction behavior, we logged all viewport scrolls and feature usage during the reading session along with timestamps. This allows us to measure the frequency of feature usage of TermSight and the baseline interface.

6.4 Analysis

In this section, we introduce our analysis framework for the user study. We used a causal framework to understand the effects of the treatment and a Bayesian analysis to estimate the treatment effect (§6.4.1). Then, we discuss the thematic analysis (§6.4.2) used to analyze the qualitative responses from the semi-structured interviews.

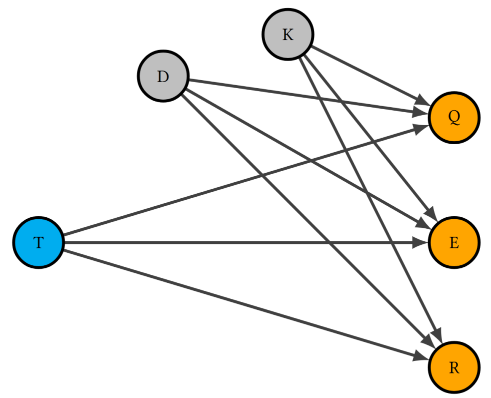


Figure 6: The causal graph (DAG) has three types of nodes: treatment (T); pre-treatment covariates D (demographics), K (knowledge of Terms of Service); and measured outcomes Q (comprehension), R (recall), and E (user experience). Since we use randomized assignment for the interface type (treatment vs. control), ToS type (social media vs. e-commerce), and presentation order (first vs. second), the treatment T is independent of the pre-treatment covariates D and K. To estimate the causal effect of the treatment T on the outcomes Q, E, and R, we can use a simple regression model, without conditioning on covariates D and K, since there is no backdoor path from the treatment T to the outcome variables.

6.4.1 Bayesian Analysis. First, we used a structural causal framework popularized by Pearl [84] to understand the effects of the treatment (Figure 6 shows the DAG for the study). More details about structural causal framework can be found in Appendix F.1. Next, we used Bayesian inference to estimate the treatment effect of TermSight on the outcomes of interest. While the use of Bayesian estimates is growing in HCI [53, 58], we briefly justify its use over traditional non-Bayesian methods. First, as Kay et al. [53] points out, a Bayesian framework leads to an accumulation of knowledge within HCI, where the posterior of the parameters in a prior work can serve as the prior in the current experiment. Second, a Bayesian model is transparent—the researcher will foreground all the assumptions in the analysis via their model. Third, use of a Bayesian framework shifts the discussion from "did it work" to "effect size of the intervention" [53]. While NHST techniques can compute the confidence interval, these intervals are susceptible to misinterpretation [11, 45] and, importantly, underemphasized in favor of p-values. Finally, as McElreath [71] points out, a Bayesian model with the use of maximum entropy priors for the parameters (e.g., the normal distribution) is *the most conservative* given the evidence to estimate the effect of the treatment.

6.4.2 Thematic Analysis. To analyze the qualitative responses from the semi-structured interviews, we adopted the six phases of reflexive thematic analysis suggested by Braun and Clarke [19], following the same procedure as our formative study (§3.1.2).

7 Findings

7.1 RQ1: How did participants perceive TermSight and its features?

17 participants highlighted that the baseline interface was easier to read than most of the ToS they had seen in the past, which often lacked the top navigation panel or a table of contents. When asked about which version of the interface they prefer, 19 participants preferred TermSight. One participant preferred the baseline interface, citing difficulty navigating TermSight's two scrollable columns on a small screen with a trackpad that lacks scroll functionality. Below we present our quantitative findings on the experience outcomes (§7.1.1), followed by qualitative findings organized around recurring themes about features of TermSight (§7.1.2 – §7.1.5).

7.1.1 TermSight improved all seven user experience measures. Figure 8 shows participants' ratings of the 7 reading experience measures for both interfaces. Our Bayesian analysis (model details in Appendix F.2) reveals a significant effect of the TermSight interface on the reading experience measures. Participants found reading ToS with TermSight to require significantly less effort. When navigating ToS, deciding which policy to read and what text to read within a policy was significantly easier with TermSight. Moreover, participants were significantly more willing to read ToS and spend more time reading ToS with TermSight. The forest plots shown in Appendix Figure 14 demonstrate that the 94% Highest Posterior Density Interval (HPDI[3]) for each treatment–control pair for every user experience measure does not overlap with each other. This indicates a significant positive effect of the treatment on all 7 user experience measures.

In Figure 7, we show the contrast between the treatment and control conditions averaged across 7 user experience measures as well as the posterior of the effect size. The posterior distribution of the effect size (Cohen's d) of the treatment on user experience is centered around 2.7, indicating a large effect size[4]. Because the HPDI doesn't overlap with the ROPE[5], the effect of TermSight on user experience is significant. To further disentangle how individual features of TermSight contributed to the improvement in user

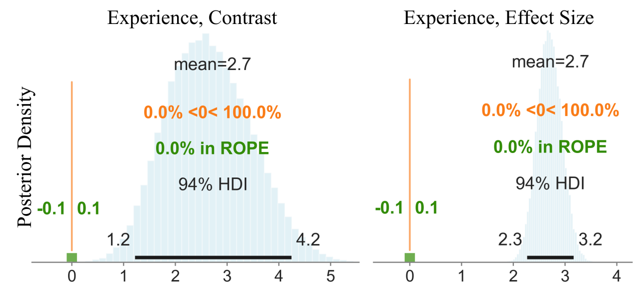


Figure 7: The posterior distribution of the contrast between treatment and control on user experience is shifted to the right, with the posterior distribution of the effect size centered around 2.7 without overlapping with ROPE, suggesting a significant and large effect size.

experience, we present recurring themes related to participants' perceptions and feedback on features of TermSight.

7.1.2 Power Meter reduced navigation cost. All 20 participants described that the Power Meter helped them gain an intuition of the content within each policy and supported decision-making on which policy to pay attention to.

> *"The color system was really guiding a lot of my decisions on which policies to read. The vibrant red was always what I try to get to first and the things that I cared the most about. So, finding which terms I'm agreeing to that I have the least amount of agency."* (P12)

P6 used the TermSight system first. However, he lamented the fact that when using the baseline interface without the Power Meter, deciding which policy to read becomes a guessing game.

> *"Policies like Terms of Service or privacy policy are kind of broad. Does that apply to something I would actually care about? I don't know. So I'd have to click each one and have to go through it. It was more like a guessing game."* (P6)

Additionally, eight participants described how the document preview reduced navigation cost by allowing them to access information from other policies without *"clicking on the link and possibly losing where I am on"* (P1).

7.1.3 Summary Snippets and Color provided guidance and helped surface power and relevance within documents. Eighteen participants described how the color of the Summary Snippets helped surface where power lies within a document, which is challenging to "see" otherwise.

> *"The color coding gave me a better understanding of who's benefiting more, how it's benefiting them, and how it affects me. Versus [in the baseline], it feels like it only really benefits the company itself."* (P18)

Moreover, all participants described how the color and summary allowed them to make decisions and prioritize what to focus on strategically.

[3]In Bayesian analysis, the 94% HPDI refers to that smallest interval of the posterior distribution that has 94% of the probability mass. It is common in Bayesian analysis to use intervals distinct (97% and 89% HPDI intervals are also used) from the typical 95% value to avoid the confusion of the frequentist confidence intervals.
[4]This is the effect size measured not on the outcome Likert scale (1–5) but on the latent scale of the model coefficients.
[5]We use a Region of Practical Equivalence (ROPE) of [-0.1, 0.1]. ROPE is a region where the difference between the treatment and control conditions is considered practically equivalent. This prevents us from focusing on small differences that are not practically meaningful.

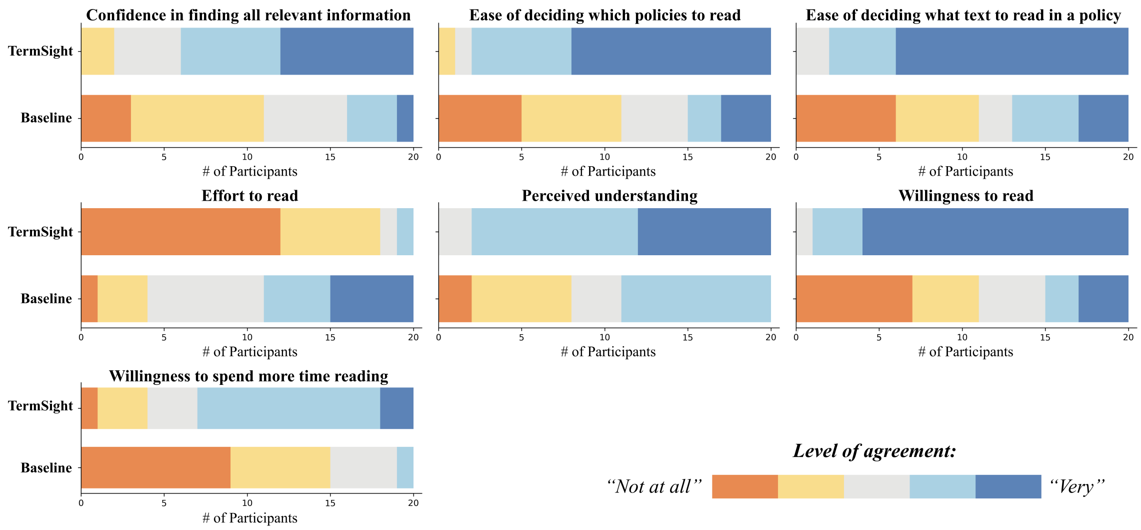


Figure 8: Participants' ratings of reading experiences. Participants found it to be easier to find and understand relevant information with TermSight. Moreover, participants were more willing to read and spend more time on ToS with TermSight.

> *"I really liked having the summary with colors. I had a hierarchy of how I was going to read stuff. I made sure I read all the [saturated] red ones, and then the [saturated] green. If I had time, I would skim the [saturated] yellow and see if anything stuck out. It would take me a lot longer to filter for information otherwise."* (P6)

While participants described focusing on the more saturated colors, especially red and green, six participants complimented how TermSight made AI decisions (i.e., classification of power and relevance) transparent by keeping all the Summary Snippets visible and described skimming the less saturated ones to validate and make sure they didn't miss information.

7.1.4 Summary Snippets simplified and broke down dense text to support sensemaking. Summary Snippets *"broke down the entire section into a few bullet-pointed summaries"* (P11), which helped participants surface and absorb relevant information hidden in dense text (20/20). Nine participants further described how the Summary Snippets simplified the text and served as a scaffold to encourage reading the original text.

> *"Not only does it [TermSight] bring me up to things that I would normally miss. but then that would be an introduction or a guide, or a push towards reading the whole 3 paragraphs. ... It is really guiding me through this entire document, and it's simplifying it."* (P2)

P20 described the experience of using Summary Snippets to make sense of the big ideas of the section while being able to dive deeper as forming *"a web of thought and ideas: what everything kind of means together and piece by piece"*. Not only do the Summary Snippets serve as an entry point to the original text, four participants explicitly highlighted how the Summary Snippets helped them check their understanding after reading the original text.

7.1.5 Term definitions and scenarios helped consume original text and envision implications. Fourteen participants noted that the system identified phrases attracted their attention and prompted them to check on their understanding of the phrase that they may not have noticed otherwise. Moreover, the definitions and scenarios made legal language more approachable for participants (14/20).

> *"It really just made it a lot easier to understand, because most of the time they write it in legal terms that most people don't know. People don't speak legal language. So this broke it down easier for most people like the Layperson."* (P17)

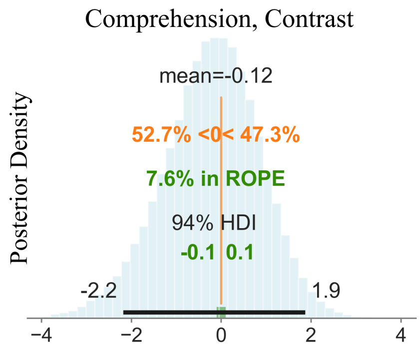
12 participants highlighted how the scenarios helped envision potential implications of the phrase in the context of the user's intended usage of the service and was *"easier to resonate"* (P17).

> *"I like that it provides a real-world example of when and how [the phrase] would be relevant to someone. I think it actually does better explaining the concept than the definition. Because it's not just the definition. It's the definition in context."* (P11)

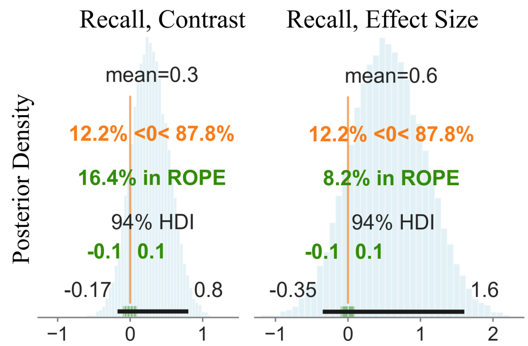
Moreover, one participant (P1) noted how the scenarios helped her come up with more questions, which she asked with the Ask function. Through this process, TermSight helped P1 to quickly obtain information from related policies.

7.2 RQ2: How did TermSight influence comprehension and recall?

7.2.1 Comprehension. Out of 6 questions, participants on average scored 2.00 (σ = 1.03) when using TermSight and 2.15 (σ = 1.23) when using the baseline interface. Our Bayesian analysis for the comprehension quiz (model details in Appendix F.3) reveals no significant differences in the comprehension scores across interface conditions. As shown in Figure 9a, the posterior distribution of the contrast between treatment and control is centered around zero.



(a) Posterior distribution of contrast (between treatment and control) on comprehension.



(b) Posterior distribution of contrast (between treatment and control) and effect size on recall.

Figure 9: In (a), the distribution centers around zero and overlaps with ROPE, indicating no effect of treatment on comprehension. (b) illustrates a positive influence of treatment on recall with the posterior distribution of the effect size centered around 0.6, suggesting a medium effect size. Because the HPDI overlaps with ROPE, there is no significant effect of treatment on recall.

When the question ID and service type are fixed, there are significant overlaps between the 94% HPDI for each treatment–control pair as shown by the forest plots in Appendix Figure 15, suggesting no significant differences. While we aimed to design comprehension questions with one single answer based on the original text, we acknowledge that legal contracts and problems may contain room for interpretation [68]. Therefore, we performed a supplemental, more conservative analysis by removing any question where the applicability of the original text to the question may contain room for interpretation, potentially leading to different answer choices (removed 2 out of 6 for each service type). The findings remained consistent, showing no significant difference in comprehension. More details of the analysis can be found in Appendix G.

7.2.2 Recall. Participants on average recalled 1.54 (σ = 1.96) correct facts using TermSight and 1.30 (σ = 1.72) correct facts using the baseline interface. Our Bayesian analysis for the recall task (model details in Appendix F.4) reveals no significant differences in the recall scores across interface conditions. Figure 9b shows the posterior distribution of the effect size of the treatment on recall (μ = 0.6, HPDI = [-0.35, 1.6]). Despite the mean being 0.6, suggesting a medium effect size, the HPDI interval has an overlap with the ROPE, implying no significant effect of the treatment on recall.

7.3 RQ3: How did participants read with TermSight and its features?

7.3.1 Feature Usage. In the baseline interface, participants on average navigated to 7 policies (σ = 5). Participants described trying to skim everything from top to bottom and often skipped difficult sections. 11 participants clicked the section heading in the table of contents at least once to navigate to targeted sections.

As shown in Figure 10 and 11, participants used most features of TermSight throughout the session. When using TermSight, participants on average navigated to 6 policies (σ = 3). 19 participants hovered over a Power Meter to gain an overview of documents in ToS for more than 1 second. 13% (17/135) of the total usage was for previewing documents hyperlinked inline, demonstrating that the Power Meter provided value also for inline hyperlinks.

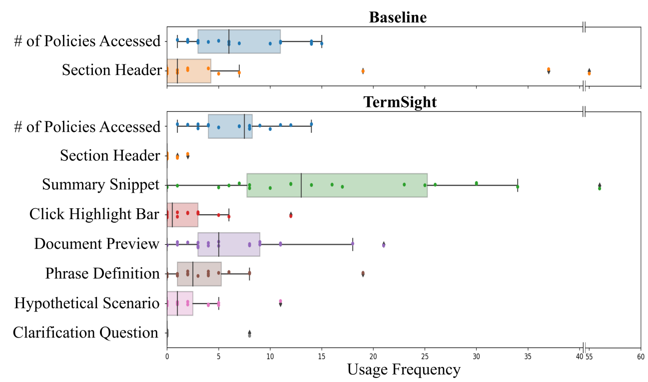


Figure 10: Feature usage. Each dot represents the number of times a feature is used by one participant during a reading session. Participants used most features of TermSight. In contrast to the baseline, participants, when using TermSight, more frequently clicked the Summary Snippets as opposed to the section headers to navigate to the original text.

Summary Snippets were the most frequently used feature. 19 participants clicked the Summary Snippets to navigate and read the original text at least once (μ = 17, σ = 13). In contrast, participants rarely clicked the section headers to navigate to the original text (4 clicks in total across 3 participants). This demonstrates that Summary Snippets were potentially more useful as navigational features to surface and interrogate relevant information from the original text compared to section headers. In addition, 10 participants clicked the Highlight Bar to read the Summary Snippet after reading the original text to check their understanding.

For Phrase Scope, 16 participants accessed phrase definitions and 12 participants viewed hypothetical scenarios. Out of 77 instances when Phrase Scope was accessed, 72 were suggested by TermSight. This demonstrates the value of suggesting difficult or ambiguous phrases as opposed to solely relying on users to notice them.

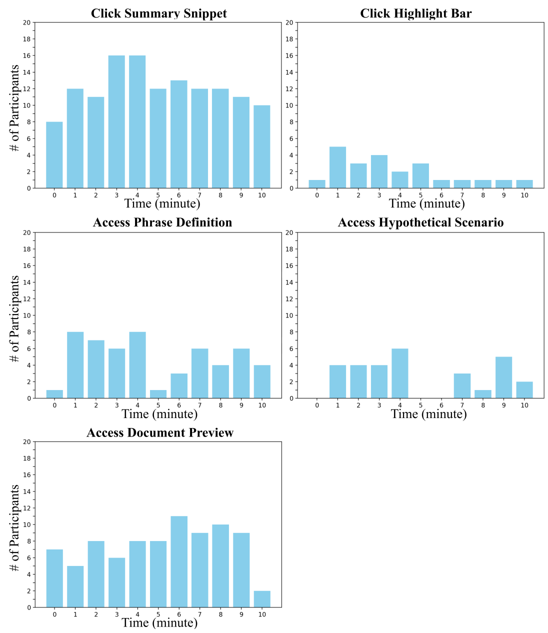


Figure 11: Minute-by-minute usage of features of TermSight during the ten-minute reading task. Participants used these features throughout the reading session, rather than being limited to the beginning or end. The Summary Snippet is the most frequently used feature throughout.

7.3.2 Reading Strategies for Accessing Original Text. We observed diverse ways in which participants used TermSight to read ToS (Figure 12). Interestingly, these reading strategies centered around the use of the original text. In some cases, participants used TermSight's features as *entries into* the original text, while in others, the features were used as *verification of* the original text. In still others, participants used features in tandem with the text to *calibrate trust* in TermSight. Below, we provide case studies of these strategies.

Summary driven reading. 19 out of 20 participants at times relied on the Summary Snippets as a primary entry point for engaging with the text. One notable example of this behavior was P3, who mainly read the Summary Snippets and clicked 11 of them to interrogate the original text (Figure 12a). P3 described how colors were driving his attention to read. P3's decision to click the Summary Snippets and interrogate the original text relied on the color coding and evaluation of the information loss of the Summary Snippets.

> *"I mainly stay on the left side. I read them if it was like a saturated red or green. I clicked on them because I felt like there was more information for me [behind the Summary], and I want to learn more about it."* (P3)

Original text driven reading. On the other hand, nine participants went through at least one policy focusing mainly on the original text. These participants described how they wanted to first read the actual text before relying on AI-generated summaries due to worries about the potential imperfections of AI. When reading the privacy policy, P8 focused on the original text as shown by the frequent scrolling of the original text in Figure 12b. In addition, P8 clicked the Highlight Bar 11 times to refer back to the Summary Snippets. P8 described how he knew AI was not perfect from his prior professional experiences in AI. As a result, P8 took the colors and Summary Snippets as a guide rather than a guarantee.

> *"I'm not gonna just go and trust it without doing my due diligence. I would take it as a guide. I won't take it as a guarantee, because AI is not perfect. I think it [AI] would try to label it the best it can and I think it did, to be honest."* (P8)

Calibration of trust. Six participants described that they went through a calibration process with AI-generated features such as the Summary Snippets and color coding. For example, P6 described how his reading behavior was different during the calibration process and after. At the beginning of the reading session, P6 focused on reading the original text to evaluate the colors and summaries (original text driven reading). Afterward, P6 relied on the summaries a lot more by scrolling and skimming the left panel and clicking on the Summary Snippets to dive to the original text, as seen in Figure 12c (summary driven reading).

> *"Early on in the reading, I read the whole thing on the right first. Then, I looked at the summaries, and once I realized that it did do a good job summarizing it with colors. I trusted it enough for the rest of the time."* (P6)

8 Discussion

In this paper, we explored how AI-powered systems might support ToS contract reading given the unique challenges that legal text poses (i.e., value-dependent, ambiguous, and legally binding). Our formative study revealed the barriers ToS presents, which informed the design of an augmented reading interface, TermSight. Results from our user study (N=20) suggest that TermSight's features made ToS contracts easier to read and navigate, without an increase or decrease in comprehension. Below, we discuss the implications of our findings around the barriers ToS contracts present (§8.1), the design implications for future contract reading interfaces (§8.2), and the limitations of technology in mitigating the societal difficulties with service contracts (§8.3).

8.1 From Simplifying Legal Text to Navigating Dark Patterns

Past work on augmenting legal text has focused on simplification (e.g., [67]). In the context of ToS contracts, our formative study revealed that the barriers people face extend well beyond complex text. Participants struggled to navigate multiple policies within a ToS, surface relevant information from dense legal language, and interpret ambiguous terminology. Surprisingly, existing navigational affordances—such as policy names, section headers, or overview summaries—often obfuscated rather than clarified the original text, a design reminiscent of dark patterns in the HCI literature [16, 59, 69]. These barriers may not be unique to ToS, but reflect broader challenges for reading legal contracts taken to their extreme. For example, leases, insurance policies, and employment

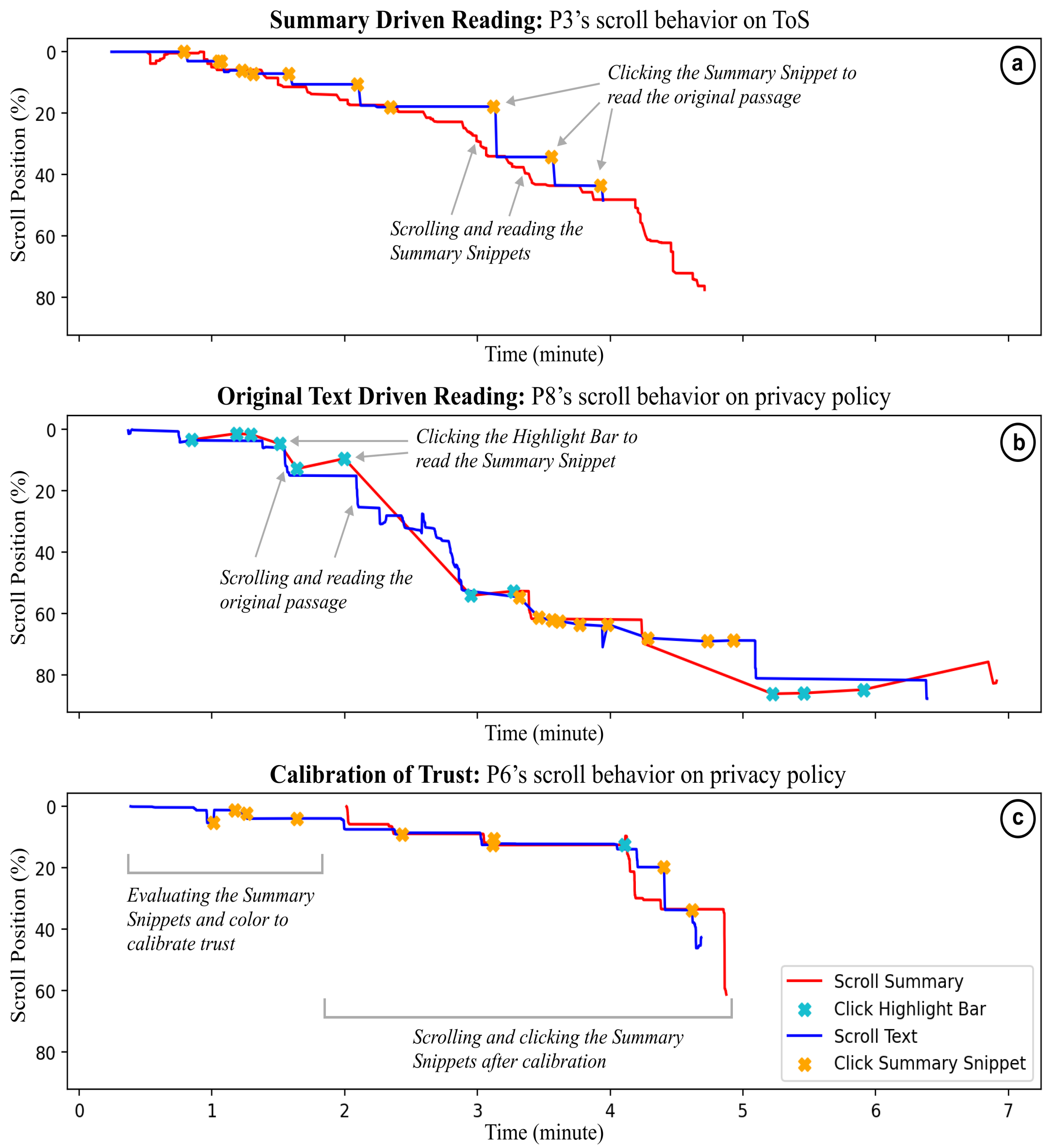


Figure 12: Scrolling behavior of participants who employed each non-exclusive reading strategy. TermSight features two scrollable columns: Summary Snippets on the left (red lines) and Original Text on the right (blue lines). Clicking a Summary Snippet (orange X) auto-scrolls the Original Text panel to the referenced text, while clicking the Highlight Bar (light blue X) auto-scrolls the Summary panel to the corresponding Summary Snippet. (a) Summary-driven reading occurs when participants primarily read and scroll the Summary Snippets, occasionally clicking on them to interrogate the original text (19/20). (b) Original text-driven reading occurs when participants first read the original text and use the Summary Snippets as a supplement (9/20). (c) The calibration of trust mainly occurs at the beginning of the session, where participants would compare and contrast the original text to the Summary Snippets with color (6/20).

contracts similarly feature multiple addenda; long, dense text; and ambiguous language that obscures consumer rights [33, 74]. Future work could expand our findings by conducting large-scale analyses of dark patterns in contracts, an area largely underexplored by the HCI community. In addition, future contract reading interfaces should consider going beyond simplifying legal text to helping users navigate contracts while combating the dark patterns of design. TermSight serves as an initial exploration for inspiring future contract reading tools in this direction.

8.2 Design Implications and Opportunities

TermSight is a preliminary exploration of designing contract reading interfaces. Below, we discuss the design implications of our findings and opportunities for further exploration.

8.2.1 ***Providing guidance at multiple levels of granularity***. While prior work focuses on simplifying legal text (e.g., [67]), TermSight demonstrated the potential values of guiding readers at the contract-level (e.g., Power Meter and document previews), document-level (e.g., color-coded Summary Snippets), and phrase-level (e.g., phrase identification). Participants in our user study affirmed these designs in both self-report and interaction behaviors (§7). For example, participants reported how Power Meter transformed deciding which policy to read from a "guessing game" to a more deliberate decision-making process. On the other hand, Summary Snippets with color and Phrase Scope drew participants' attention to snippets of information and phrases buried in dense text that they might otherwise miss. Consequently, finding which policy to read and what text to read within a policy was significantly easier for participants with TermSight. The specificity of TermSight's guidance (i.e., relevance classification) relies on the user persona. For the user study, we designed two personas to be able to control for their quality (§6.1.2). These personas also aligned with participants' profiles in the user study (Figure 13). Future work could explore allowing users to further customize their personas to receive personalized guidance and overcome the barriers of doing so (e.g., forming information bubbles [91]).

8.2.2 ***Prompting reflection on unintended consequences***. Prior work has explored providing term definitions to aid reading academic papers [8, 44]. However, legal contracts are unique because contracts are agreements for the unforeseen future rather than a description of the past [42]. Reflecting this characteristic of contract, participants in our formative study mentioned that definitions alone may not always be sufficient for legal language because it can be difficult to contextualize abstract phrases in unforeseen future scenarios. In the user study, participants described the value of Phrase Scope for presenting scenarios to prompt reflection on possible implications of signing a contract, similar to how prior work has encouraged scientists to reflect on the unintended consequences of their research [83, 104]. Some participants even noted how the scenario was more helpful than the definition (§7.1.5). Future work could further explore how to design these scenarios and investigate their influence on user perception of the contractual clauses, such as by manipulating the linguistic framing [90] or showing information from legal cases and news articles.

8.2.3 ***Centering augmentations around the original text***. Legal contracts differ fundamentally from documents like academic papers or news articles because the language itself carries legal power [22]. Consequently, LM-powered text transformations cannot substitute the original text. TermSight explored the approach of tightly linking generated text with the original text through visual placement (i.e, side-by-side views) and deep linking (i.e., linking to narrowly attributed text like a single clause), allowing readers to directly compare language and navigate to the original text in at most one click. When given these features, participants took advantage of them to engage with the original text (§7.3.2). Some participants dove into the original text by finding relevant Summary Snippets, while others verified their understanding of the original text with the Summary Snippets. Some participants even explicitly evaluated the Summary Snippets with the original text to calibrate trust in the system. An exciting avenue for future work would be to go beyond supporting original text reading to encouraging it. Features that nudge readers toward the original text where appropriate, such as revealing summaries only after reading [21], are an interesting start, but little work has evaluated these features' effects in situations where the original text might be discouraging to engage with.

8.3 Technological Limitations and Legal Implications

Compared to simplified reading tasks (e.g., reading a single clause [62, 88] or policy overview [57, 97]), our user study involved realistic ToS with over 10 policies. TermSight made ToS significantly easier to read and navigate. However, TermSight did not worsen or improve comprehension, reminiscent of prior findings that simplified language often did not improve comprehension of privacy policies [57, 88, 97]. There may be multiple explanations behind this finding. For example, Summary Snippets may have distracted attention away from the original text, the 12-word limit of Summary Snippets may have constrained the conceptual coverage of the original text, and participants' calibration behaviors (i.e., confirming summary information with the original text) may have distracted them from the reading task. However, our findings may have also revealed the broader limitations of technological solutions in mitigating the societal difficulties with service contracts such as ToS. Although TermSight offered multi-level guidance on relevance (e.g., Power Meter, Summary Snippets), the sheer volume of potentially relevant information in ToS remains a dominant challenge. For example, in the social media ToS, 172/759 (23%) discrete Information Snippets were classified by TermSight as being relevant to the given user persona (§6.1.2), including 6 snippets needed to answer the comprehension questions. Within this subset, being able to notice and recall snippets of information to correctly answer the comprehension questions may be difficult. While future deployment of TermSight could allow users to customize the personas for more targeted guidance, this may risk introducing information bubbles [4, 24, 34, 91]. Given that the formation of a binding contract under U.S. law is premised on the doctrine of the 'duty to read'[6], our findings raise fundamental questions about the practicality of this duty for ToS. Instead of relying solely on technological fixes, policy

[6]Contracting parties are legally responsible for reading and understanding the contract before signing [12].

and legal innovations are needed. For example, policy makers and legal scholars could explore alternative paradigms of contracting that operationalize ToS into shorter, context-specific contracts to reduce the upfront burden to consent.

8.4 Limitations and Future Work

The nature of this work is exploratory. Similar to prior HCI studies (e.g., [20, 21, 51]), we designed and studied TermSight to explore the challenges and opportunities of designing contract reading interfaces, rather than proposing TermSight as the definitive solution. Future work can pursue finer-grained exploration through controlled ablation studies. Examples may include investigating the influence of Power Meter on ToS reading frequency in the wild, assessing how hypothetical scenarios affect user perception of clauses (e.g., trust [93]), or investigating how AI affordances (e.g., Summary Snippets) support or distract from original text reading (e.g, [76]) through eye-tracking.

Because participants were paid to read ToS in a timed reading task, their motivation and behavior may not reflect real-world usage. Moreover, a within-subject design may lead to ordering effects or fatigue. In the study, we counterbalanced the conditions and didn't observe a significant effect of ordering on user experience, comprehension, and recall (Appendix Figure 16). Yet, it may still have influences on participants' interaction. Future work can expand our study by conducting between-subject or field experiments.

In addition, our participants were recruited from Prolific, mostly college-educated, and fluent in English, which may limit the generalization of our findings to the broader population. Marginalized communities, including neurodivergent readers and individuals with limited English literacy, can face additional barriers when interpreting legal documents. For example, readers with ADHD might have a hard time staying focused when reading visually dense text such as ToS [13, 95]. Future work can investigate how to help alternative reader populations make sense of legal contracts by evaluating TermSight with these populations of users or integrating additional supports (e.g., translation).

Future systems could adapt our design of TermSight to other contracts. For example, Summary Snippets may help readers surface and interpret relevant or predatory clauses in leases, while Phrase Scope may help patients envision the unintended consequences of signing a medical contract through patient-specific scenarios. Finally, features of TermSight rely on the capabilities of LMs, which may produce imperfect outputs [46, 66, 70]. While work is actively investigating how to guide LMs to generate factually correct information [103], we believe a validation mechanism for LLM output is necessary for future deployment of systems like TermSight, such as by facilitating end-user [49, 60] or expert auditing of AI output [50].

9 Conclusion

We regularly sign legal contracts for where we live, who we work with, and how we interact with digital services. Yet, these contracts have drifted from their premise of facilitating mutual understanding to complex documents that discourage reading. In this work, we explored the opportunities and limitations of intelligent reading support for legal contracts, using ToS contracts as a case study. Our formative study revealed ineffective and deceptive designs at all levels of a ToS. To make service contracts approachable, we designed and evaluated an intelligent reading interface: TermSight. Participants reported that TermSight reduced the difficulty of engaging with ToS and increased their willingness to do so. Additionally, features of TermSight enabled participants the ability to surface and approach relevant information at all levels of contract reading. Taken together, TermSight presents one avenue for making legal contracts more approachable to the general public. However, facilitating a deeper understanding of ToS remains an open challenge due to the overwhelming amount of information in ToS.

References

[1] 2024. *PrivacySpy.* Retrieved Jan 10, 2025 from https://privacyspy.org
[2] 2024. *Terms of service; didn't read.* Retrieved April 2, 2024 from https://tosdr.org/
[3] 2025. *Reddit User Agreement.* Retrieved August 8th, 2025 from https://redditinc.com/policies/user-agreement
[4] Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (Chicago, Illinois) *(LinkKDD '05)*. Association for Computing Machinery, New York, NY, USA, 36–43. doi:10.1145/1134271.1134277
[5] Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 743–749. doi:10.18653/v1/2020.findings-emnlp.66
[6] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2165–2176. doi:10.1145/3442381.3450048
[7] Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know Your Audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 14, 26 pages. doi:10.1145/3613904.3642289
[8] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–38.
[9] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies* 43, 1 (2014), 1–35.
[10] Benjamin B Bederson, James D Hollan, Allison Druin, Jason Stewart, David Rogers, and David Proft. 1996. Local tools: An alternative to tool palettes. In *Proceedings of the 9th annual ACM symposium on User interface software and technology.* 169–170.
[11] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods* 10, 4 (2005), 389.
[12] Uri Benoliel and Shmuel I Becher. 2019. The duty to read the unreadable. *BCL Rev.* 60 (2019), 2255.
[13] Barbara Bental and Emanuel Tirosh. 2007. The relationship between attention, executive functions and reading domain abilities in attention deficit hyperactivity disorder and reading disorder: A comparative study. *Journal of Child Psychology and Psychiatry* 48, 5 (2007), 455–463.
[14] Jaspreet Bhatia, Travis D Breaux, Joel R Reidenberg, and Thomas B Norton. 2016. A theory of vagueness and privacy risk perception. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*. IEEE, 26–35.
[15] Eric A Bier, Maureen C Stone, Ken Pier, William Buxton, and Tony D DeRose. 1993. Toolglass and magic lenses: the see-through interface. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques.* 73–80.
[16] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies* (2016).
[17] Daniel Braun and Florian Matthes. 2021. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Proceedings of the 1st Workshop on NLP for Positive Impact.* 93–99.
[18] Daniel Braun, Elena Scepankova, Patrick Holl, and Florian Matthes. 2017. SaToS: Assessing and Summarising Terms of Services from German Webshops. In

Proceedings of the 10th International Conference on Natural Language Generation, Jose M. Alonso, Alberto Bugarín, and Ehud Reiter (Eds.). Association for Computational Linguistics, Santiago de Compostela, Spain, 223–227. doi:10.18653/v1/W17-3534

[19] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[20] Yining Cao, Jane L E, Chen Zhu-Tian, and Haijun Xia. 2023. DataParticles: Block-based and language-oriented authoring of animated unit visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[21] Xiang "Anthony" Chen, Chien-Sheng Wu, Lidiya Murakhovs' ka, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2023. Marvista: exploring the design of a human-AI collaborative news reading tool. *ACM Transactions on Computer-Human Interaction* 30, 6 (2023), 1–27.

[22] Mindy Chen-Wishart. 2012. *Contract law*. Oxford University Press, USA.

[23] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (A) I am not a lawyer, but...: engaging legal experts towards responsible LLM policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2454–2469.

[24] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2021. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (Aug. 2021), 89–96. doi:10.1609/icwsm.v5i1.14126

[25] Lorrie Faith Cranor, Serge Egelman, Steve Sheng, Aleecia M McDonald, and Abdur Chowdhury. 2008. P3P deployment on websites. *Electronic Commerce Research and Applications* 7, 3 (2008), 274–293.

[26] Lorrie Faith Cranor, Praveen Guduru, and Manjula Arjula. 2006. User interfaces for privacy agents. *ACM Transactions on Computer-Human Interaction (TOCHI)* 13, 2 (2006), 135–178.

[27] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for? An eye-tracking study of information usage in web search. In *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems*. 407–416.

[28] Robert Lee Dickens. 2007. Finding common ground in the world of electronic contracts: the consistency of legal reasoning in clickwrap cases. *Marq. Intell. Prop. L. Rev.* 11 (2007), 379.

[29] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. 2016. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1450–1461.

[30] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2024. Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–21.

[31] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 476–490. doi:10.1145/3581641.3584034

[32] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[33] Meirav Furth-Matzkin. 2017. On the unexpected use of unenforceable contract terms: Evidence from the residential rental market. *Journal of Legal Analysis* 9, 1 (2017), 1–49.

[34] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.

[35] Nathaniel Good, Rachna Dhamija, Jens Grossklags, David Thaw, Steven Aronowitz, Deirdre Mulligan, and Joseph Konstan. 2005. Stopping spyware at the gate: a user study of privacy, notice and spyware. In *Proceedings of the 2005 symposium on Usable privacy and security*. 43–52.

[36] Nathaniel S Good, Jens Grossklags, Deirdre K Mulligan, and Joseph A Konstan. 2007. Noticing notice: a large-scale experiment on the timing of software license agreements. In *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems*. 607–616.

[37] Alfonso Guarino, Nicola Lettieri, Delfina Malandrino, and Rocco Zaccagnino. 2021. A machine learning-based approach to identify unlawful practices in online terms of service: analysis, implementation and evaluation. *Neural Computing and Applications* 33 (2021), 17569–17587.

[38] Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Lu Wang, and Tal August. 2024. Personalized jargon identification for enhanced interdisciplinary communication. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, Vol. 2024. 4535.

[39] Helena Haapio and Stefania Passera. 2017. Contracts as interfaces: exploring visual representation patterns in contract design. *Legal Informatics, Cambridge, UK: Cambridge University Press. Published ahead of print as part of doctoral dissertation* 37 (2017).

[40] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. 2021. Toggles, dollar signs, and triangles: How to (in) effectively convey privacy choices with icons and link texts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–25.

[41] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 531–548.

[42] Oliver Hart. 2017. Incomplete contracts and control. *American Economic Review* 107, 7 (2017), 1731–1752.

[43] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[44] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[45] Rink Hoekstra, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review* 21, 5 (2014), 1157–1164.

[46] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2024).

[47] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. 2023. Memory sandbox: Transparent and interactive memory management for conversational agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[48] Duha Ibdah, Nada Lachtar, Satya Meenakshi Raparthi, and Anys Bacha. 2021. "Why Should I Read the Privacy Policy, I Just Need the Service": A Study on Attitudes and Perceptions Toward Privacy Policies. *IEEE access* 9 (2021), 166465–166487.

[49] Farnaz Jahanbakhsh, Amy X Zhang, Karrie Karahalios, and David R Karger. 2022. Our Browser Extension Lets Readers Change the Headlines on News Articles, and You Won't Believe What They Did! *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–33.

[50] Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, et al. 2024. Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7194–7219.

[51] Hita Kambhamettu, Danaë Metaxa, Kevin Johnson, and Andrew Head. 2024. Explainable notes: Examining how to unlock meaning in medical notes with interactivity and artificial intelligence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[52] Robin Bradley Kar and Margaret Jane Radin. 2019. Pseudo-contract and shared meaning analysis. *Harvard Law Review* 132, 4 (2019), 1135–1219.

[53] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (Santa Clara, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 4521–4532. doi:10.1145/2858036.2858465

[54] Matthew Kay and Michael Terry. 2010. Textured agreements: re-envisioning electronic consent. In *Proceedings of the sixth symposium on usable privacy and security*. 1–13.

[55] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. doi:10.1145/1572532.1572538

[56] Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasarathy. 2020. Toward Domain-Guided Controllable Summarization of Privacy Policies.. In *NLLP@ KDD*. 18–24.

[57] Jana Korunovska, Bernadette Kamleitner, and Sarah Spiekermann. 2020. The Challenges and Impact of Privacy Policy Comprehension. ECIS.

[58] Vinay Koshy, Tanvi Bajpai, Eshwar Chandrasekharan, Hari Sundaram, and Karrie Karahalios. 2023. Measuring User-Moderator Alignment on r/ChangeMyView. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 286 (oct 2023), 36 pages. doi:10.1145/3610077

[59] Lin Kyi, Sushil Ammanaghatta Shivakumar, Cristiana Teixeira Santos, Franziska Roesner, Frederike Zufall, and Asia J Biega. 2023. Investigating deceptive design in GDPR's legitimate interest. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[60] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34.
[61] Logan Lebanoff and Fei Liu. 2018. Automatic Detection of Vague Words and Sentences in Privacy Policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3508–3517.
[62] Mina Lee, Jake M Hofman, David M Rothschild, and Daniel G Goldstein. 2025. Could AI Make Legalese Comprehensible to the Public? *Available at SSRN* (2025).
[63] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* 27 (2019), 117–139.
[64] Xingyu Liu, Annabel Sun, and Jason I Hong. 2021. Identifying Terms and Conditions Important to Consumers using Crowdsourcing. *arXiv preprint arXiv:2111.12182* (2021).
[65] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael J. Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2024. The Semantic Reader Project. *Commun. ACM* 67, 10 (Sept. 2024), 50–61. doi:10.1145/3659096
[66] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. Factual consistency evaluation of summarization in the Era of large language models. *Expert Systems with Applications* 254 (2024), 124456.
[67] Laura Manor and Junyi Jessy Li. 2019. Plain English Summarization of Contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*. 1–11.
[68] Florencia Marotta-Wurgler and David Stein. 2025. Building a Long Text Privacy Policy Corpus with Multi-Class Labels. *NYU Law and Economics Research Paper Forthcoming* (2025).
[69] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.
[70] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1906–1919.
[71] Richard McElreath. 2015. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
[72] Microsoft. 12/15/2016. *P3P is no longer supported*. https://learn.microsoft.com/en-us/previous-versions/windows/internet-explorer/ie-developer/compatibility/mt146424(v=vs.85)
[73] George R Milne and Mary J Culnan. 2004. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of interactive marketing* 18, 3 (2004), 15–29.
[74] Warren Mueller. 1970. Residential tenants and their leases: An empirical study. *Michigan Law Review* 69, 2 (1970), 247–298.
[75] Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9612–9631. doi:10.18653/v1/2024.emnlp-main.538
[76] Aileen Nielsen, Stavroula Skylaki, Milda Norkute, and Alexander Stremitzer. 2023. Effects of XAI on Legal Process. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (Braga, Portugal) *(ICAIL '23)*. Association for Computing Machinery, New York, NY, USA, 442–446. doi:10.1145/3594536.3595128
[77] Razieh Nokhbeh Zaeem, Ahmad Ahbab, Josh Bestor, Hussam H. Djadi, Sunny Kharel, Victor Lai, Nick Wang, and K. Suzanne Barber. 2022. PrivacyCheck v3: Empowering Users with Higher-Level Understanding of Privacy Policies. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1593–1596. doi:10.1145/3488560.3502184
[78] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K. Suzanne Barber. 2020. PrivacyCheck v2: A Tool that Recaps Privacy Policies for You. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3441–3444. doi:10.1145/3340531.3417469
[79] Patricia A Norberg, Daniel R Horne, and David A Horne. 2007. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of consumer affairs* 41, 1 (2007), 100–126.
[80] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.
[81] Przemyslaw Palka. 2023. Terms of injustice. *W. Va. L. Rev.* 126 (2023), 133.
[82] Shidong Pan, Thong Hoang, Dawen Zhang, Zhenchang Xing, Xiwei Xu, Qinghua Lu, and Mark Staples. 2023. Toward the cure of privacy policy reading phobia: Automated generation of privacy nutrition labels from privacy policies. *arXiv preprint arXiv:2306.10923* (2023).
[83] Rock Yuren Pang, Sebastin Santy, René Just, and Katharina Reinecke. 2024. Blip: facilitating the exploration of undesirable consequences of digital technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
[84] Judea Pearl. 2009. *Causality*. Cambridge university press.
[85] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.
[86] Joseph Reagle and Lorrie Faith Cranor. 1999. The platform for privacy preferences. *Commun. ACM* 42, 2 (1999), 48–55.
[87] Daniel Reinhardt, Johannes Borchard, and Jörn Hurtienne. 2021. Visual interactive privacy policy: The better choice?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
[88] Eric P Robinson and Yicheng Zhu. 2020. Beyond "I agree": Users' understanding of web site terms of service. *Social media+ society* 6, 1 (2020), 2056305119897321.
[89] Abhilasha Sancheti, Aparna Garimella, Balaji Srinivasan, and Rachel Rudinger. 2023. What to Read in a Contract? Party-Specific Summarization of Legal Obligations, Entitlements, and Prohibitions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14708–14725. doi:10.18653/v1/2023.emnlp-main.909
[90] Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9977–10000.
[91] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[92] Donghoon Shin, Lucy Lu Wang, and Gary Hsieh. 2024. From paper to card: transforming design implications with generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.
[93] Ben Shneiderman. 2000. Designing trust into online experiences. *Commun. ACM* 43, 12 (2000), 57–59.
[94] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6829–6839.
[95] Pnina Stern and Lilach Shalev. 2013. The role of sustained attention and display medium in reading comprehension among adolescents with ADHD and without it. *Research in developmental disabilities* 34, 1 (2013), 431–439.
[96] Simon Sweeney and Fabio Crestani. 2006. Effective search results summary size and device screen size: Is there a relationship? *Information processing & management* 42, 4 (2006), 1056–1074.
[97] Madiha Tabassum, Abdulmajeed Alqhatani, Marran Aldossari, and Heather Richter Lipford. 2018. Increasing user attention with a comic-based policy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[98] Lee Taber, Paul May, Keane Yahn-Krafft, and Steve Whittaker. 2020. Beyond avoidance and passivity: Novel uis to make terms of service comprehensible. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.
[99] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. 2021. Defining privacy: How users interpret technical terms in privacy policies. *Proceedings on Privacy Enhancing Technologies* (2021).
[100] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. I read but don't agree: Privacy policy benchmarking using machine learning and the EU GDPR. In *Companion Proceedings of the The Web Conference 2018*. 163–166.
[101] Noriko Tomuro, Steven Lytinen, and Kurt Hornsburg. 2016. Automatic summarization of privacy policies using ensemble learning. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*. 133–135.
[102] Kim-Phuong L Vu, Vanessa Chambers, Fredrick P Garcia, Beth Creekmur, John Sulaitis, Deborah Nelson, Russell Pierce, and Robert W Proctor. 2007. How users

read and comprehend privacy policies. In *Human Interface and the Management of Information. Interacting in Information Environments: Symposium on Human Interface 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part II*. Springer, 802–811.
[103] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2024. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*. 13697–13720.
[104] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–40.
[105] Sewall Wright. 1934. The Method of Path Coefficients. *The Annals of Mathematical Statistics* 5, 3 (1934), 161–215. http://www.jstor.org/stable/2957502
[106] Haijun Xia, Bruno Araujo, Tovi Grossman, and Daniel Wigdor. 2016. Object-oriented drawing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4610–4621.
[107] Johnathan Yerby and Ian Vaughn. 2022. Deliberately confusing language in terms of service and privacy policy agreements. *Issues in Information Systems* 23, 2 (2022).
[108] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 1–18.
[109] Grace Q Zhang. 2015. *Elastic language: How and why we stretch our words*. Cambridge University Press.
[110] Tiancheng Zhao and Kyusong Lee. 2020. Talk to Papers: Bringing Neural Question Answering to Academic Search. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 30–36.

A Iterative Design Study

A total of 8 participants were recruited through Prolific to evaluate an early prototype of TermSight. This prototype closely resembled the current version of TermSight but lacked a document preview when hovering over the Power Meter. Additionally, the prototype used three saturation levels for relevance, resulting in a total of nine colors (3 hues for power x 3 saturation levels for relevance). Participants could also toggle between two layouts for the Summary Snippets: condensed layout and in-context layout. The condensed layout matched the current design shown in Figure 1. In contrast, the in-context layout placed each Summary Snippet directly next to its corresponding Information Snippet. After a brief interface tutorial, participants were given 10 minutes to read a ToS using the interface, followed by a semi-structured interview about their experience. Each session lasted for 40 minutes. Below, we present the main findings and changes made.

Participants preferred condensed layout for quick navigation. 7 participants preferred and primarily used the condensed layout of the Summary Snippets during the reading session. They noted that the purpose of the Summary Snippets was to support navigation, and the condensed layout made it easier to gain an overview and identify relevant information in a policy. In contrast, the in-context layout required significantly more scrolling, which participants found overwhelming. As a result, in the final version of TermSight, we used the condensed layout (Figure 1).

Reduce visual complexity of the color scheme. All participants found the colors to be intuitive and effective in highlighting power and relevance, helping them decide what to read. However, 3 participants pointed out that while the power dimension was intuitive, the relevance dimension with 3 saturation levels was challenging to differentiate, as there would be 9 colors in total. To reduce visual complexity, the final version of TermSight included only two saturation levels (High vs. Low) for relevance (Figure 2).

Complement Power Meter with Document Preview. Additionally, 4 participants suggested that while the Power Meter provided a general overview, they wanted a more concrete preview. This feedback led to the integration of a document preview feature in the final version of TermSight, which appears when users hover over the Power Meter (Figure 3).

B Additional Implementation Details

B.1 Source Document and Pre-processing

Given a source file in HTML or markdown, the document is first segmented by headers (e.g., h1, h2, h3, h4). Within each section or subsection, the text is further chunked by newline separators (i.e., "\n") into segments of around 1,500 characters (approximately 250 words) using langchain's RecursiveCharacterTextSplitter[7], with no overlap between chunks. Importantly, paragraph structures are preserved, as the text splitter only splits at newline breaks.

B.2 Obtaining Summary Snippets and Information Snippets

The prompt used to obtain the Summary Snippets and Information Snippets is detailed in Figure 17. We constrained the Summary Snippets to be short because prior works have found that adding short summaries (10-20 words) under search results was more effective for navigational and information-seeking tasks compared to longer summaries, which can be harder to skim [27, 96].

B.3 Classifying Information Snippets

In TermSight, two classifications were performed for each Information Snippet using GPT-4o with few-shot prompting (Figure 18 and 19). For the classification of Power, each Information Snippet was classified by the degree of control or agency it grants to the Service Provider or the User (Categories: Service, Neutral, User). For the classification of Relevance, Information Snippets were classified based on their relevance to the user persona (Categories: High, Low). User personas used for the study are included in Appendix E.2.

B.4 Phrase Scope

B.4.1 Identifying unfamiliar or ambiguous phrases. For each document chunk, we employ two few-shot prompts to GPT-4o to identify potentially unfamiliar (Figure 20) or vague phrases (Figure 21). Few-shot examples for identifying unfamiliar and ambiguous phrases were selected based on phrases participants found challenging in our formative study. Both prompts were applied to each document chunk, producing two sets of identified phrases. Since a phrase may be both unfamiliar and vague, the union of these two sets was taken as the final set of identified phrases for a given chunk of text.

B.4.2 Generating phrase definitions and answers to user questions. We first retrieve potentially relevant document chunks from the vector database. The database query is framed as a question: *"What does {input phrase} refer to in the sentence: {phrase context}"*. Here, phrase context refers to the chunk of text containing the input phrase. Then, the query is embedded using OpenAI's text-embedding-3-small model, and the top 15 chunks from the vector database are

[7] https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/

retrieved based on cosine similarity. Figure 22 shows the prompt used to generate the definitions after retrieving potentially relevant document chunks. When users ask additional questions, the same retrieval-augmented question answering pipeline used for generating definitions is applied, with the only difference being the question asked (Figure 24).

B.4.3 Generating scenarios. We leveraged GPT-4o to generate customized scenarios of potential implications based on user persona. The prompt is specified in Figure 23.

C Evaluation of TermSight Output

C.1 Classification of Information Snippets

Out of the sampled 116 snippets, our evaluation revealed 3 instances where the power classifications were imperfect, mainly caused by the lack of context in the input Information Snippet. For example, the snippet (*'All Buy Now purchases in a ServiceY Show are final and binding'*) was classified by TermSight as favoring the service. In the generated explanation, the assumption was that the statement meant that the user can not return or refund the purchase. However, users can return and get a refund if there are problems with the purchase (e.g., an item doesn't match its description) as stated in the service's return policy, making this snippet more neutral. The return policy was not included in the snippet context; as a result, this neutral statement was considered more service-oriented. We didn't observe any clause favoring the service provider being misclassified as neutral or favoring the user.

For the classification of relevance, there were 2 instances, out of 116, where irrelevant Information Snippets were classified as relevant to the persona. In both cases, the snippets were more relevant to sellers, not buyers, even though the persona given was that of a buyer. There were 11 instances where the Information Snippets classified as relevant were indirectly relevant to the input user personas. For example, sellers have to pay a fee to the platform after selling an item. Despite this information being more relevant to sellers, TermSight classified it as relevant to buyers, as it exposed the potential hidden fees that sellers might include as part of their listing price. We show examples of imperfect classification of Power and Relevance in Table 1.

C.2 Term Definitions and Scenarios

Our evaluation revealed that all the generated definitions were correct, but out of the 113 definitions, 4 were only general definitions of the phrase, not specific to the ToS. On further inspection, we identified that this was because these vague phrases were not explicitly defined anywhere in the ToS. Additionally, their meanings are service-specific and cannot be extrapolated from common sense or inferred by large language models (LLMs). For example, for the phrase *'aggregated anonymized statistics'*, TermSight provided a general definition of what it might mean to aggregate and anonymize user data. However, the specific details—such as what user data is being aggregated—were not specified in the ToS. Rather than a limitation of TermSight, these imperfections highlight the ill-defined nature of the language used in ToS.

For the generated scenarios of phrases relevant to the user persona, we found one instance where the scenario was factually incorrect based on the input context. The input context and phrase stated that users do not gain ownership rights by downloading content from the service. However, the scenario claimed that users might lose ownership rights over the content they create by uploading it to the service. We also noticed that when the passages or phrases target a different audience (e.g., developers) than the user persona used to generate the scenarios (i.e., a lay user), the generated scenarios become less relevant or useful. We did not regenerate these imperfections to keep the user experience realistic in real-world settings where the LLM outputs are not guaranteed to be perfect. Participants in our user study were informed that AI output can be imperfect. This specific incorrect scenario was not accessed by any participants in the study.

D Study Materials: Formative Study

Questions asked before the reading session:

- How familiar are you with Terms of Service in general?
- Have you previously read or wished that you read the Terms of Service? What were your reasons for wanting or not wanting to read the Terms of Service?
- Have you used your assigned service before?
- Have you read the ToS for your assigned service before?

Semi-structured interview questions asked after the reading session:

- What were the challenges you faced when reading the Terms of Service?
- How did you go about reading the ToS?
- What information were you interested in in ToS? Both from your prior experience in reading ToS and this ToS?
- Imagine if you have a magic wand that can transform the ToS in whatever ways you want. How would you transform the Terms of Service?

E Study Materials: User Study

E.1 Baseline Interface

The baseline ToS reading interface used in the user study can be found in Figure 25.

E.2 User Persona

Two personas were given to the participants during the user study for the social media service (content consumer who posts personal content) and the e-commerce service (buyer who rarely posts reviews). The same personas were used for features of TermSight to classify relevance and generate personalized scenarios. The personas were designed based on information that participants in the formative study described caring about.

Classification	Input Snippet	Output	Imperfection
Power	"All Buy Now purchases in a ServiceY Show are final and binding."	Service	Lack of enough context in the input. Users do have the option to return. It's a neutral clause.
Power	"Use public content for any illegal, deceptive, unethical, false, misleading, or improper purpose, including the infringement of third-party intellectual property rights."	Neutral	Lack of enough context in the input. The clause describes what third party licensee cannot do with user content. It's a user-benefiting clause.
Relevance	"Your earnings are based on the listing price and actual earnings will vary based on the final order price, Seller discounts, and any other applicable taxes and discounts."	High	Lack of enough context in the input. This clause is for sellers and is less relevant to the input buyer persona.
Relevance	"ServiceY cannot guarantee that a ServiceY consignment listing will be sold or that a certain sales amount will be earned for individual items or an entire shipment."	High	Lack of enough context in the input. This clause is for sellers and is less relevant to the input buyer persona.

Table 1: Examples of imperfect classifications of Power and Relevance.

Persona: Content consumer who posts personal content
Imagine you are a lay user of social media platforms. You are over 18 years old and located in the United States.
Your usage of Social Media sites:

- You spend most of your time on the platform scrolling through feeds, liking posts, chatting with other users, and sharing personal content such as photos.

Things you care about when using Social Media Sites:

- You care about Privacy, particularly what data is being collected and how your data can be used and shared.
- You care about what the service can do with user-generated content, such as licenses over user content or advertising with user content.
- You care about potential liabilities when using ServiceX.

Persona: Buyer who rarely posts reviews
Imagine you are a lay user of E-commerce platforms. You are over 18 years old and located in the United States.
Your usage of E-commerce sites:

- You typically engage with the E-commerce platform to buy new or used items from other users.
- You rarely post any reviews or content on the service.

Things you care about when using E-commerce sites:

- You care about information related to making purchases, refunds, returns, user protection policies, termination, arbitration, and liabilities.
- You also care about Privacy, particularly what data is being collected and how your data can be used and shared.

E.3 Participant Demographics

Of the 20 participants, 13 self-identify as female and 7 as male. 1 participant had a high school degree, 3 had an associate degree, 12 had a bachelor's degree, and 4 had a master's degree. 4 participants were between the ages of 18-25, 5 were between 26-35, 5 were between 36-45, and 6 were between 46-55. When asked about how many ToS they've read, 4 participants read none, 5 read between 1–3, 5 read between 4–6, 1 read between 7–9, and 5 read greater than 10. Participants in the user study also found their given user persona for both services to highly align with their personal usage of the service and their personal value (Figure 13).

E.4 Pre-survey Questions

Questions that were asked once at the beginning of the interview:

- For how many online platforms have you read their Terms of Service (ToS) before? [None (0), Few (1-3), Some (4-6), Many (7-9), A lot (>10)]
- How familiar are you with Terms of Service (ToS) for online platforms? (5-point Likert rating)

5-point Likert rating questions that were asked before each of the two reading sessions:

- How familiar are you with (E-commerce or Social Media) sites?
- How well does the above user persona align with your personal usage of (E-commerce or Social Media) sites?
- How well does the above user persona align with things you personally care about when using (E-commerce or Social Media) sites?

E.5 5-point Likert Ratings of Reading Experience

Participants rated their reading experience after each of the two reading sessions.

- **E1:** I'm interested in spending more time reading the service's Terms of Service (ToS) with the current interface and wish to get a link after the study.
- **E2:** How hard did you have to work to read the ToS?
- **E3:** How easy was it for you to decide which sub-policies to read?
- **E4:** How easy was it for you to decide what text to read within a sub-policy?
- **E5:** How confident are you that you got all the relevant information from the ToS (including sub-policies)?
- **E6:** How much do you feel like you understood the ToS?
- **E7:** How much would you be willing to read the ToS of other services with this interface?

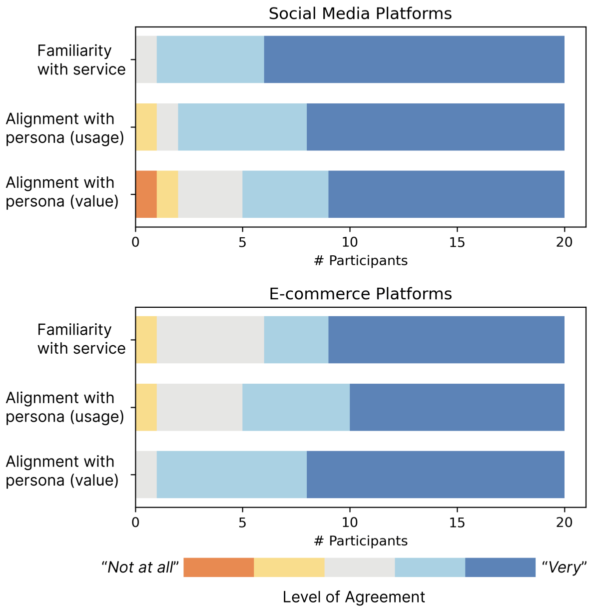


Figure 13: Participants' self-ratings of their familiarity with social media and e-commerce platforms and their alignment with the user persona given for each service. Most participants were familiar with both social media and e-commerce platforms and found the given persona to align with their personal usage of the service and personal value.

E.6 Semi-structured Interview Questions

Below are questions that were asked after each of the two reading sessions, except the last question. The last question was only asked after the second reading session, where participants would compare and contrast both interfaces.

- Describe your experience reading the ToS using the interface.
- How did you read the ToS using the interface?
- What were the challenges you faced when reading the Terms of Service? To what extent did the interface help you? (Be specific about the features in the interface).
- Which interface do you prefer? Why? Compare and contrast. (Asked only after the second reading session)

F Bayesian Analysis Model Details and Additional Analysis

In this section, we provide details on the Bayesian models. We first explain the concepts of structural causal framework and DAG in §F.1. All models were informed by the DAG in Figure 6. Next, in §F.2, we describe the model used to estimate the effect of the treatment on the user experience. In §F.3, we describe the model used to estimate the effect of the treatment on the comprehension outcomes, i.e., the quiz questions. Finally, in §F.4, we describe the model used to estimate the effect of the treatment on the experiment subjects' ability to recall facts correctly. Figure 14 and 15 show the forest plots of model coefficients for the experience and comprehension outcomes. Figure 16 presents forest plots illustrating the effect of ordering on user experience, comprehension, and recall.

F.1 Structural Causal Framework

We used a structural causal framework popularized by Pearl [84] to better understand the effects of the treatment. A structural causal model framework comprises exogenous variables U, endogenous variables V, and a set of functions F that represent the physical process that results in the values of the endogenous variables. For example, the question $v_i = f(v, u)$ implies that the values of v and u completely determine the value of v_i through the function f. The structural causal model can be represented as a Directed Acyclic Graph (DAG) where the nodes represent the variables and the edges represent the causal relationships among the variables. Thus, in the preceding example, v and u are the parents of v_i in the DAG. The structural causal model framework is non-parametric, but frequently used with generalized linear models to estimate the functions f. The structural causal framework has connections with structural equation models [105] used in the Economic Sciences, though the latter were developed with linear relationships in mind. Figure 6 shows the DAG for the study.

F.2 Modeling Experience Outcomes

We asked the experiment subjects to rate their experience of using the treatment or the baseline interface on a Likert scale from 1 to 5 (L=5 outcome values). There were N subjects, and we asked each subject K (K=7) questions. Since the Likert scale is ordinal, we modeled the experience outcomes using an ordinal regression model. The DAG in Figure 6 indicates that the experimental condition alone affects the experience outcomes. There were three randomizations: whether the subject was shown the treatment or control interface (i.e., interface type), whether the subject was shown the social media ToS or the e-commerce ToS (i.e., service type), and whether the treatment was shown first or second (i.e., presentation order). We modeled the experience outcomes using a cumulative `logit` model with a linear link function:

$$Y_{i,j,k,l,m} \sim \text{OrderedLogistic}(\kappa, \phi_{i,j,k,l,m}) \tag{1}$$
$$\phi_{i,j,k,l,m} = \alpha_i + c_{j,k} + s_l + o_m \tag{2}$$

Below, we specify the priors for α_i (intercept, one per participant), $c_{j,k}$ (joint variable for the interface type and user experience item), s_l (service type), o_m (presentation order), and κ_i (cutpoints).

$$\alpha_i \sim \text{Normal}(0, 1), i \in \{1, \ldots, N\} \tag{3}$$
$$c_{j,k} \sim \text{Normal}(0, 1), j \in \{1, 2\}, k \in \{1, \ldots, K\} \tag{4}$$
$$s_l \sim \text{Normal}(0, 1), l \in \{1, 2\} \tag{5}$$
$$o_m \sim \text{Normal}(0, 1), m \in \{1, 2\} \tag{6}$$
$$\kappa_i \sim \text{Normal}(0, 1), i \in \{1, \ldots, L-1\} \tag{7}$$

In eq (7), we further ensured that the samples are ordered, i.e., $\kappa_1 < \kappa_2 < \cdots < \kappa_{L-1}$. The priors are conservative. For example, on the logit scale, a coefficient that is normally distributed with mean 0 and standard deviation 1, the range [-3, 3] covers 99% of the distribution and evaluates using inverse logistic to the outcome probability range of [0.04, 0.95].

F.3 Modeling Comprehension Outcomes

There were N subjects, and we asked each subject K (K=6) questions to test their comprehension. The outcome is a binary variable with 1 indicating a correct answer and 0 indicating an incorrect answer. The DAG in Figure 6 indicates that the experimental condition alone affects the comprehension outcomes. There were three randomizations: whether the subject was shown the treatment or control interface, whether the subject was shown the social media ToS or the e-commerce ToS, and whether the treatment was shown first or second. We modeled the comprehension outcomes using a logistic regression model. Notice that we created a joint variable $c_{j,k,l}$ for the coefficients corresponding to the interface type, question, and service type. We did this because the comprehension questions were different across the two services used in the experiment. We included the interface variable (i.e., treatment vs. control) to be able to easily estimate the effect of the treatment on the comprehension outcomes per question. The model is given by:

$$Y_{i,j,k,l,m} \sim \text{Binomial}(p_{i,j,k,l,m}) \quad (8)$$

$$\text{Logistic}(p_{i,j,k,l,m}) = \alpha_i + c_{j,k,l} + o_m \quad (9)$$

Below, we specify the priors for α_i (intercept, one per participant), $c_{j,k,l}$ (joint variable for the interface type, comprehension question, and service type), and o_m (presentation order).

$$\alpha_i \sim \text{Normal}(0, 1), i \in \{1, \ldots, N\} \quad (10)$$

$$c_{j,k,l} \sim \text{Normal}(0, 1), j \in \{1, 2\}, k \in \{1, \ldots, K\}, l \in \{1, 2\} \quad (11)$$

$$o_m \sim \text{Normal}(0, 1), m \in \{1, 2\} \quad (12)$$

As in the previous model, the priors are conservative. For example, on the logit scale, a coefficient that is normally distributed with mean 0 and standard deviation 1, the range [-3, 3] covers 99% of the distribution and evaluates using inverse logistic to the outcome probability range of [0.04, 0.95].

F.4 Modeling Recall Outcomes

There were N subjects, and we asked each subject to recall facts from the ToS. The outcome that we measure is the number of correctly recalled facts, making the outcome a count variable. Since many of the subjects could not recall any facts correctly, we modeled the recall outcomes using a Zero Inflated Poisson regression model. The DAG in Figure 6 indicates that the experimental condition alone affects the comprehension outcomes. As before, there were three randomizations: whether the subject was shown the treatment or control interface, whether the subject was shown the social media ToS or the shopping ToS, and whether the treatment was shown first or second. The model is given by:

$$Y_{i,j,k,l} \sim \text{ZeroInflatedPoisson}(\lambda_{i,j,k,l}, \phi) \quad (13)$$

$$\text{Log}(\lambda_{i,j,k,l}) = \alpha_i + c_j + s_k + o_l \quad (14)$$

Below, we specify the priors for α_i (intercept, one per participant), c_j (interface type), s_k (service type), o_l (presentation order), and ϕ (probability of zero inflation).

$$\alpha_i \sim \text{Normal}(0, 1), i \in \{1, \ldots, N\} \quad (15)$$

$$c_j \sim \text{Normal}(0, 1), j \in \{1, 2\} \quad (16)$$

$$s_k \sim \text{Normal}(0, 1), k \in \{1, 2\} \quad (17)$$

$$o_l \sim \text{Normal}(0, 1), l \in \{1, 2\} \quad (18)$$

$$\phi \sim \text{Beta}(2, 2) \quad (19)$$

As in the previous model, the priors are conservative. For example, on the Log scale, a coefficient that is normally distributed with mean 0 and standard deviation 1, the range [-3, 3] covers 99% of the distribution and evaluates using inverse log to the outcome range of [0.05, 21]. Notice that we model the number of correctly recalled facts, and thus is a conservative prior. Given that most subjects could not recall any facts, the zero inflation parameter is set to a weakly informative prior using a Beta distribution.

G Supplemental Analysis for Comprehension Outcomes

We aimed to design the comprehension questions to have a single answer that best matches the clauses in the original text. However, we acknowledge that legal problems and contractual clauses may leave room for interpretation [68]. As a result, we run an additional, more conservative analysis by removing any question that may contain room for interpretation on whether the original text would apply to the situation being asked in the question, leading to different answer choices. We removed 2 out of 6 questions for each service type (Social Media: Q1, Q4; Shopping: Q2, Q3). For all other questions, there is a single correct answer choice directly matching the original text, while all the other options either conflict with or were never mentioned in the original text. Below is one example of the questions removed from this analysis:

Social Media Q1: How can ServiceX use photographs you post for advertisements or promotions?

Original Answer: ServiceX can use photos you post in ads without your permission and is not obligated to attribute you as the creator.

Relevant Clause: "...you grant us a worldwide, royalty-free, perpetual, irrevocable, non-exclusive, transferable, and sublicensable license to use, copy, modify, adapt, prepare derivative works of, distribute, store, perform, and display Your Content and any name, username, voice, or likeness provided in connection with Your Content in all media formats and channels now known or later developed anywhere in the world...you irrevocably waive any claims and assertions of moral rights or attribution with respect to Your Content." (*User Agreement*)

Possible Room for Interpretation: Though the original answer best matches the relevant clause. The clause does not explicitly mention advertisement. This may leave room for interpretation on whether the clause applies to ads.

We use the same Bayesian model as §F.3, except that there are 4 comprehension questions for each service that are analyzed (i.e., K=4). Out of the 4 questions for each service, participants on average scored 1.5 (σ = 1.0) when using TermSight and 1.6 (σ = 1.1) when using the baseline interface. Our Bayesian model analysis reveals no significant differences in the comprehension scores across interface conditions. The posterior distribution of the contrast between treatment and control on comprehension is centered around zero and overlaps with ROPE. When the question ID and service type are fixed, there are significant overlaps between the 94% HPDI for each treatment–control pair, suggesting no significant differences for every question. These findings matches with our original findings. The figures for this additional analysis can be found in the supplemental materials (e.g., forest plots).

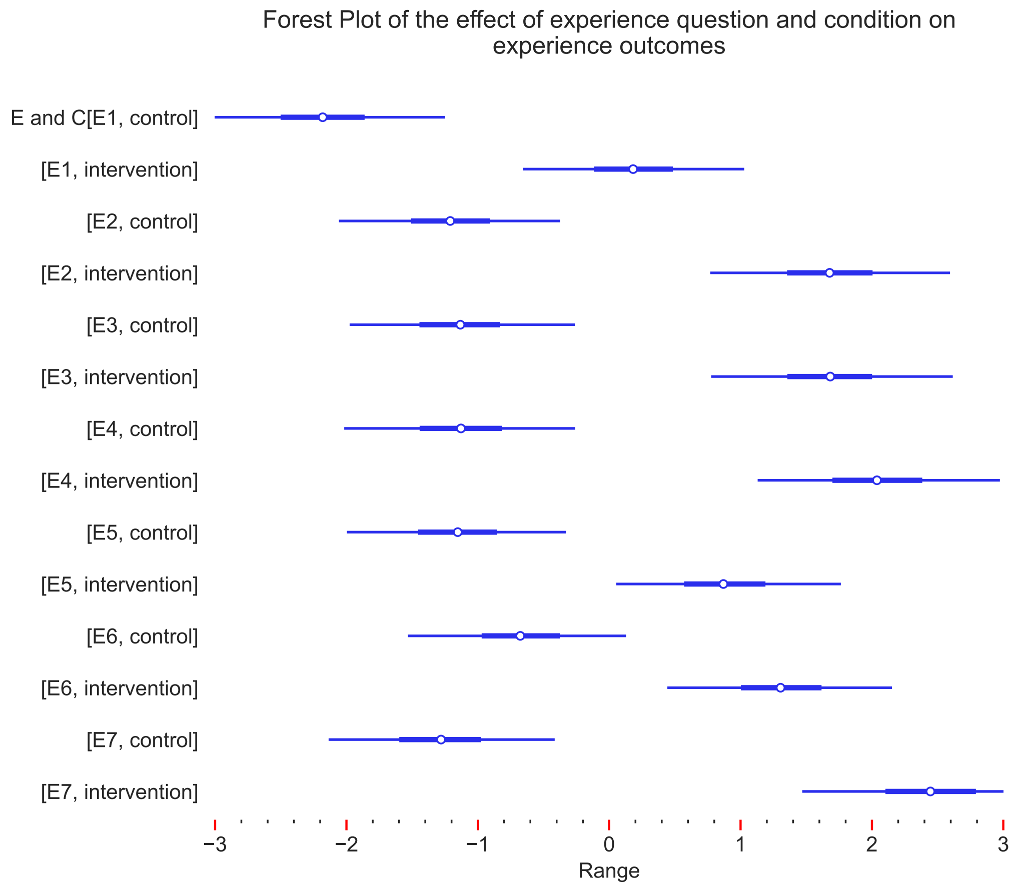


Figure 14: The figure shows a forest plot of coefficients in the model, corresponding to treatment and control, for each of the experience questions. Each line of the plot shows a 94% High Density Interval (HPDI) for the coefficient. The inner, thicker line represents the 50% HPDI. The results show a significant effect of the TermSight interface on the user experience for every question. Since the 94% HPDI for each pair (control, treatment) for every question does not overlap with each other, we should expect a significant effect of the treatment on the user experience. E1–E7 refer to the 7 experience questions specified in Appendix E.5.

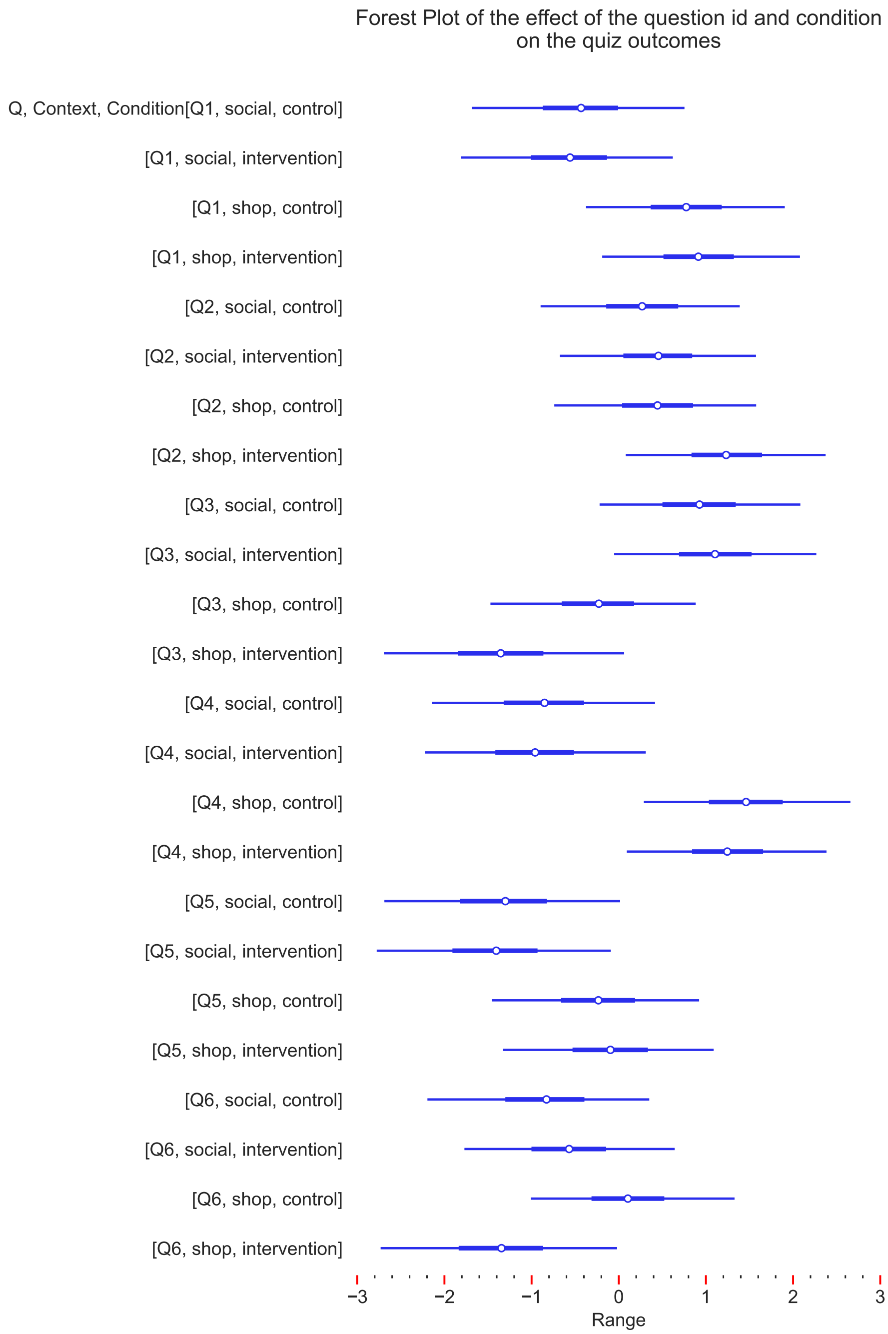


Figure 15: The figure shows a forest plot of coefficients in the model, corresponding to treatment and control, for each of the comprehension questions. The x-axis is on the logistic scale, with +3 corresponding to a 0.95 probability value on the outcome scale (-3 corresponds to 0.05). Each line of the plot shows a 94% High Density Interval (HPDI) for the coefficient. The inner, thicker line represents the 50% HPDI. The 94% HPDI intervals for each treatment–control pair, when the question and service type are fixed, show significant overlap. This suggests no significant effect of the treatment on the user's comprehension. Note that the comprehension questions Q1–Q6 are different for the two service types and are included in the supplemental materials.

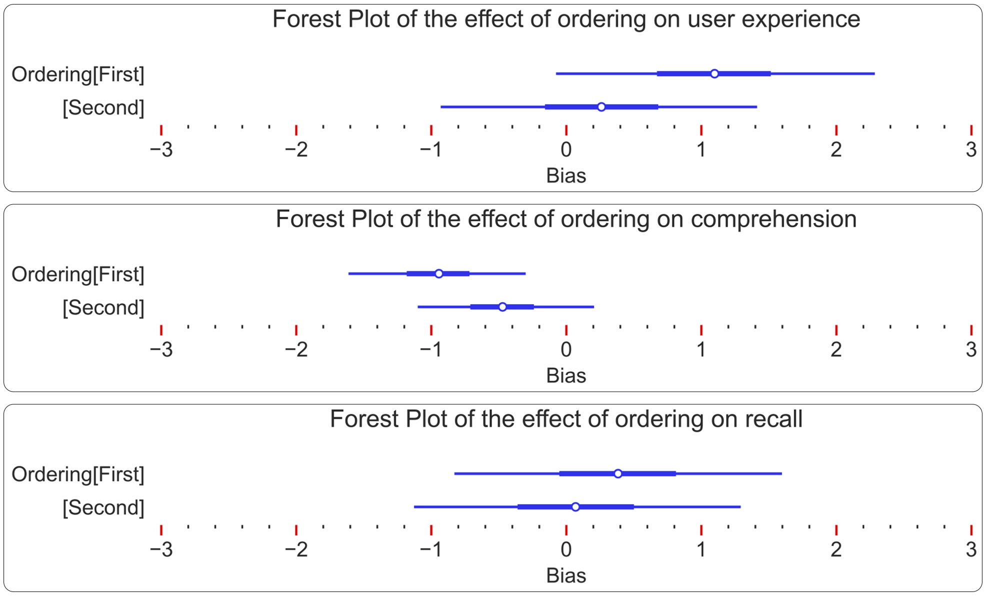


Figure 16: The figure shows the forest plots of coefficients in the model for user experience, comprehension, and recall with respect to the ordering. Each line of the plot shows a 94% High Density Interval (HPDI) for the coefficient. The inner, thicker line represents the 50% HPDI. Since the 94% HPDI for each ordering pair (first, second) overlaps, there is no significant effect of ordering on user experience, comprehension, and recall.

Prompt for Obtaining Summary Snippets and Information Snippets

Summarize the input section of the Terms of Service into concise bullet points (less than 12 words) in plain language. When adjacent paragraphs or sentences share a similar or related theme, only output 1 single bullet point. For each bullet point summary, include the full-text reference to the original passage in {} and don't use "..." to reduce text in the reference. When outputting a reference, don't change anything in the original text, such as spaces and newlines. There can be multiple sentences or paragraphs that reference a single summary. The references to summary should cover the original text.

Example output format: **{EXAMPLE OUTPUT}**

Input: **{INPUT TEXT CHUNK}**

Figure 17: Prompt for obtaining Summary Snippets and Information Snippets.

Prompt for Classifying Power

Classify the input term from a Terms of Service agreement based on the power relationship and benefit between the service and the user. Use the following categories (Service, Neutral, User):
- Service: The term grants the service provider disproportionate power or control over the user. It may impose unfair restrictions, obligations, or liabilities on the user, or reduce the user's rights and autonomy over their data or content.
- Neutral: The term outlines standard procedures, responsibilities, or conditions the user and service have. For example, users take responsibility for the content they post. It neither significantly favors the service provider nor the user, and does not substantially impact the user's rights.
- User: The term empowers the user by offering clear protections, rights, or benefits, ensuring transparency, and limiting the service provider's power.

Examples for each category:
Service:
- The service can delete specific content without prior notice and without a reason.
- The service can license user content to third parties.
- The service tracks your personal data for advertising
Neutral:
- Users are responsible for the content they post
- Users agree not to use the service for illegal purposes
- Blocking first-party cookies may limit your ability to use the service
User:
- You can opt out of targeted advertising
- The service does not sell your personal data
- The service will not allow third parties to access your personal information without a legal basis

Output format in JSON: {"Category": "Service/Neutral/User", "Explanation": "explanation of output" }

Input: **{INPUT INFORMATION SNIPPET}**

Figure 18: Prompt for classifying the power balance of the Information Snippets.

Prompt for Classifying Relevance

For the input term from a Terms of Service, output a relevance rating (High/Low) of the input term with respect to the user persona.
[High]: The term is directly relevant to the user's usage of the service or what the user cares about. The term applies to the user persona and is necessary for the user to know.
[Low]: The term is not relevant to the user's usage of the service or what the user cares about. The term doesn't apply to the user persona or is not necessary for the user to know.

User Persona: **{INPUT USER PERSONA}**

Output format in JSON: {"Relevance": "Low/High", "Explanation": "explanation of output" }

Input: **{INPUT INFORMATION SNIPPET}**

Figure 19: Prompt for classifying the relevance of Information Snippets to the user persona.

Prompt for Identifying Unfamiliar Phrases

You are a helpful assistant who extracts words or multi-word phrases in the input section of Terms of Service that a high schooler might not know the meaning of. Jargon refers to domain-specific terminologies that a lay user might not know about.

Example jargon:
- legal jargon: indemnity, arbitration, liability
- copyright licenses: sublicensable licenses, royalty-free licenses
- technical privacy terms: cookies, Ad identifiers, Authentication tokens

Return an empty array if the section does not contain jargon. The extracted word should exactly match the original input text with the same capitalization and sequence of words.

Output format in JSON: {"Jargon": []}

Input: **{INPUT TEXT CHUNK}**

Figure 20: Prompt for identifying potentially unfamiliar phrases for lay users in a text chunk.

Prompt for Identifying Vague Phrases

You are a helpful legal assistant who extracts vague terms (can have multiple words in one term) in the input section of Terms of Service. A vague term refers to information that is vaguely abstracted without a clear definition provided in the section.

Example Vague terms: information, other, some, third parties, most, generally, personal data, others, general, many, various, might, services, certain information

Return an empty array if the section does not contain vague terms. The extracted word should exactly match the original input text with the same capitalization and sequence of words.

Output format in JSON: {"Vague": []}

Input: **{INPUT TEXT CHUNK}**

Figure 21: Prompt for identifying ambiguous phrases in a text chunk.

Prompt for Generating Phrase Definitions

Use information in the retrieved context to provide a definition of the user-selected phrase or term. Avoid using long sentences. For example, if the user-selected term is "information", define what the term "information" includes and refers to, such as: location data, interaction data, profile data, etc. The output definition should be specific and straight to the point; don't include language that doesn't contribute to the definition, such as 'in the given context'. Output the string list of reference ids (["ref1", ...]) used to generate the definition under "References". If the definition of the phrase is not specified in the retrieved context, output a definition of what the phrase might mean and output an empty array for "References".

Examples: **{EXAMPLES}**

Output format in JSON: {"Definition": "", "References": ["ref1", "ref2", "ref3"]}

Retrieved Context: **{RETRIEVED CONTEXT}**

Question: What does **{INPUT PHRASE}** refer to?
Context around the user-selected phrase: **{PHRASE CONTEXT}**

Figure 22: Prompt for generating in-context phrase definitions.

Prompt for Generating Scenarios

Tell a concise what-if scenario or example in less than 50 words to demonstrate the meaning and potential implications of the user-selected phrase based on the context around the user-selected phrase. The scenario/example should be relevant to the below user persona using {an E-commerce platform of used items / a Social Media platform}.

User Persona: **{INPUT USER PERSONA}**

Output format in JSON: {"Story": ""}

User selected phrase: **{INPUT PHRASE}**
Context around the user-selected phrase: **{PHRASE CONTEXT}**
Definition of user selected phrase: **{GENERATED DEFINITION}**

Figure 23: Prompt for generating scenarios to contextualize the meaning and potential implications of a phrase.

Prompt for Generating Answers to User Questions.

You are an assistant for question-answering tasks. Use information in the retrieved context to answer the user's question in less than 5 sentences. Output the string list of reference ids (["ref1", ...]) used to generate the definition under "References". If the definition of the phrase is not specified in the retrieved context, output a definition of what the phrase might mean and output an empty array for "References".

Examples: **{EXAMPLES}**

Output format in JSON: {"Answer": "", "References": ["ref1", "ref2", "ref3"]}

Retrieved Context: **{RETRIEVED CONTEXT}**

Question: **{USER QUESTION}**
User selected phrase: **{INPUT PHRASE}**
Context around the user-selected phrase: **{PHRASE CONTEXT}**

Figure 24: Prompt for generating answers to user questions.

Phrase Type	Original Passage	Generated Definition	Generated Scenario
Jargon	"... When Your Content is created with or submitted to the Services, you grant us a worldwide, royalty-free, perpetual, irrevocable, non-exclusive, transferable, and ****sublicensable**** license to use, copy, modify, adapt, prepare derivative works of, distribute, store, perform, and display Your Content and any name, username, voice, or likeness provided ..."	Sublicensable refers to the ability to grant further licenses to third parties to use, copy, modify, or distribute your content.	Imagine posting a photo on ServiceX, and later discovering it's featured in a third-party app's advertisement. Due to the sublicensable license, ServiceX had the right to allow another company to use your photo, without needing your further permission, raising concerns about control over your content.
Jargon	"... The listed prices for Items do not include taxes, but the taxes will be displayed before a Buyer confirms the purchase. Use of ****ServiceY Credit**** (as defined below) may modify taxes that apply to a Buyer's purchase ... "	ServiceY Credit refers to non-redeemable promotional credits offered by ServiceY to be used exclusively for purchases on the Service.	Imagine Jane, a savvy shopper, receives $10 ServiceY Credit for a promotion. She buys a vintage lamp listed at $50. At checkout, ServiceY Credit reduces her total to $40, modifying applicable taxes. Jane saves money, but can't withdraw or transfer the Credit—it only applies to ServiceY purchases.
Vague	"To use certain features of our Services, you may be required to create a ServiceX account (an Account id) and provide us with a username, password, and ****certain other information**** about yourself as set forth in the Privacy Policy ..."	Certain other information refers to optional details such as a bio, gender, age, location, profile picture, or social link that you may provide when creating a ServiceX account.	Imagine signing up for ServiceX, and you're asked to provide a username, password, and 'certain other information' like your age and location. Later, you find out that ServiceX uses this data to tailor ads specifically for you, raising concerns about how much they know about you and potential privacy risks.
Vague	"... ServiceY reserves the right to discontinue providing Labels to any or all Users at any time and for ****any reason**** ..."	Any reason refers to ServiceY's discretion to stop providing shipping labels without needing to specify a particular cause or justification.	Imagine you sell vintage clothes online. ServiceY provides you with prepaid shipping labels. Suddenly, without explanation, they stop offering these labels to you. This means you'll need to cover shipping costs yourself, impacting your profits. This demonstrates ServiceY's right to discontinue services for 'any reason,' affecting your business.

Table 2: Examples of the generated definitions and scenarios for potentially unfamiliar and vague phrases.

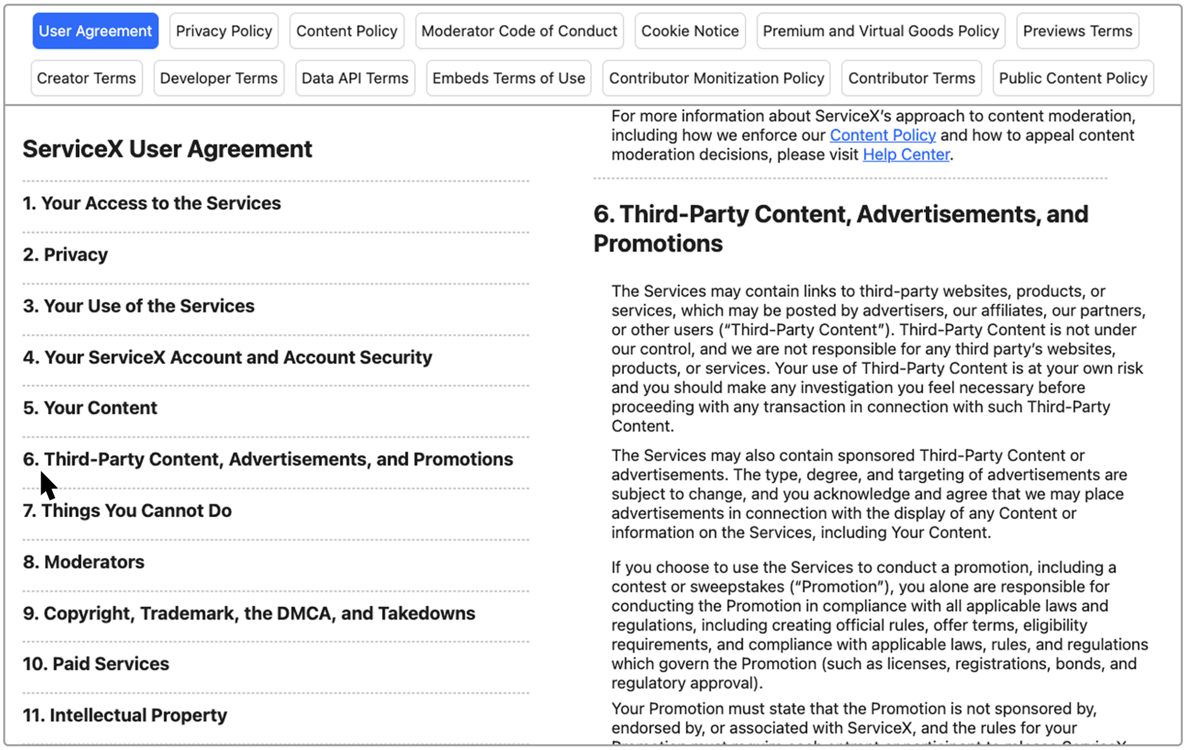


Figure 25: The baseline reading interface. Participants can navigate to different policies using the top navigational panel. In place of the Summary Snippets on the left is a table of contents. Participants can click a section header in the table of contents to navigate to the corresponding section.