プライバシーポリシーに対するユーザの理解度測定のための大規模言語モデル評価

森 啓華[1,2,a)]　伊藤 大貴[1]　福永 拓海[1]　渡邉 卓弥[1]　高田 雄太[1]　神薗 雅紀[1]　森 達哉[2,3,4]

概要：企業組織は，個人情報保護法に対応し，透明性を向上させるためにプライバシーポリシーを公表している．プライバシー情報の取り扱いとユーザ認識との間に齟齬があった場合，信頼の低下や法律違反のリスクがあるため，ユーザの理解度を測定しながらプライバシーポリシーを作成することが肝要である．しかしながら，ユーザスタディによってプライバシーポリシーを逐一評価するためには，大きな金銭的および時間的なコストを要する．本研究は，ユーザのプライバシーポリシーに対する理解度の評価を，LLM で代替可能であるかを検証する．理解を妨げる 11 種類の要素を含むプライバシーポリシーを作成し，理解度を測る設問に対するユーザと LLM の解釈を比較した．その結果，ユーザと LLM の正答率はそれぞれ平均 63.0%と 85.2%で，LLM が誤答した設問はユーザも誤答していたため，ユーザが誤解しやすい記述を LLM が検知できることが示された．また，専門用語を含むプライバシーポリシーでは，LLM だけが理解し，ユーザは理解できない傾向が見られた．これらの結果をもとに，ユーザと LLM のプライバシーポリシー理解におけるギャップを特定し，評価の自動化に向けた指針を示した．

Evaluating Large Language Models to Measure User Understanding of Privacy Policies

Keika Mori[1,2,a)]　Daiki Ito[1]　Takumi Fukunaga[1]　Takuya Watanabe[1]　Yuta Takata[1]
Masaki Kamizono[1]　Tatsuya Mori[2,3,4]

Abstract: Companies publish privacy policies to improve transparency regarding the handling of personal information. When there is a discrepancy between the description of the privacy policy and the user's understanding, it will lead to a risk of decreasing in trust. Therefore, it is essential to create a privacy policy while evaluating the user's understanding. However, periodically evaluating privacy policies through user studies requires financial and time costs. In this study, we examined whether user studies can be replaced by evaluation using LLMs. We prepared obfuscated privacy policies and questions to measure their understanding. As a result of comparing the understanding levels of users and LLMs, we found that the average correct answer rates by users and LLMs were 63.0% and 85.2%, respectively. The questions that LLMs answered incorrectly were also answered incorrectly by users. We identified the gap between users' and LLMs' understanding and provided a direction for automated evaluation of privacy policies using LLMs.

1. はじめに

プライバシーポリシーは，企業や組織が個人情報をどのように収集，使用，保護するかについての方針や手続きを公表するための主要な手段である．欧米諸国では EU 一般データ保護規則 (GDPR) やカリフォルニア州消費者プライバシー法 (CCPA) といった法制度のもとで，個人情報利用に関するユーザ伝達が不十分であった企業への制裁事例が相次ぎ [1], [2]，日本においても，2022 年 4 月の個人情報保護法改正に伴ってプライバシー遵守の基準が厳格化されている．企業組織は，ユーザからの信頼喪失や高額な制裁金のリスクを避けるために，ユーザの認識と齟齬のない

[1] デロイト トーマツ サイバー合同会社, Deloitte Tohmatsu Cyber LLC
[2] 早稲田大学, Waseda University
[3] 情報通信研究機構, NICT
[4] 理化学研究所, RIKEN
[a)] keika.mori@tohmatsu.co.jp

適切なプライバシーポリシーを作成する必要がある．

プライバシーポリシーがユーザに伝わらない主な原因は，文書としての品質と，プライバシーポリシー固有の性質に大別できる．Anca ら [3] は，文書の非論理的な構成や，二重否定などの複雑な表現が，ユーザの理解を妨げることを明らかにした．Tang ら [4] は，プライバシーポリシーに含まれる法的および技術的な専門用語によって，ドメイン知識のないユーザの理解度が低下することを指摘した．時々刻々と変容する各種サービスにおいて，個人情報の取り扱いを正確にプライバシーポリシーに反映させる必要があるため，説明の欠如や矛盾が生じてしまうことも少なくない [5], [6].

本研究が描く最終的なゴールは，ユーザにとってのプライバシーポリシーの理解しやすさを自動的に評価することである．これを達成するため，本稿では大規模言語モデル (LLM) とユーザのポリシーの理解度をそれぞれ測定する．ポリシーの理解における両者の共通点と相違点を明らかにすることで，LLM がユーザのタスクを代替できるか検証する．従来の研究 [7] では，ユーザスタディによってプライバシーポリシーを評価し，さまざまな改善点が抽出された．こうした評価を LLM によって代替できれば，時間的および金銭的コストを飛躍的に軽減させ，ユーザの理解度を高めるための評価と改善を繰り返し実施しながら，プライバシーポリシーの記述を洗練させることができる．

以上を踏まえて本研究では，**RQ1: LLM はどの程度プライバシーポリシーを理解できるか？** および **RQ2: ユーザはどの程度プライバシーポリシーを理解できるか？** に取り組む．これら RQs を解き明かすために，本研究では分析対象とするプライバシーポリシーを作成し，3 種類の LLM による理解度と，449 名に対するユーザスタディの結果を比較分析する．プライバシーポリシーには，アプリマーケットの記述に基づいてベースラインを用意した上で，ユーザが誤解しやすい 11 種類の要素を意図的に盛り込む．これらに対する理解度を測定する設問を通じて，LLM とユーザが同じように誤解する要因や，LLM だけが理解できてしまう要因を特定する．最後に，プロンプトの工夫やペルソナの設定によって LLM とユーザ間のギャップを埋め，自動的な理解度測定を実現するための方法について考察する．本研究の貢献を以下に示す．

- 3 種類の LLM および 449 名のユーザに対してアンケート形式によるプライバシーポリシーへの理解度を比較分析した結果，その正答率はそれぞれ 85.2%，63.0%となり，LLM がユーザの理解度を上回ることを示した．
- LLM とユーザ両者とも，情報の分散や欠如により，プライバシーポリシーへの理解度が下がることを明らかにした．
- ユーザは，専門知識の不足や情報の見逃しによってプライバシーポリシーへの理解度が下がるが，LLM ではこの傾向は確認されないことを明らかにした．
- 本研究で発見した LLM とユーザにおける理解度のギャップを踏まえ，LLM を用いてプライバシーポリシーの理解しやすさを評価するための指針を示した．

2. 背景および関連研究

2.1 プライバシーポリシーとユーザの理解

プライバシーポリシーは，企業組織とユーザが個人情報の取り扱いについて合意形成するためのおよそ唯一のチャネルといえる．しかし，コミュニケーション上の多くの課題が指摘されており，ユーザが読まないこと [8]，読んでも必要事項が記載されていないこと [9]，そして記載があってもユーザが理解できないこと [4], [7] が挙げられる．本研究はユーザが理解できないという課題に焦点を当て，作成したプライバシーポリシーをユーザが理解できるかどうかを自動的に評価するためのアプローチとして，LLM の応用可能性を探求する初めての研究である．

プライバシーポリシーのユーザ理解を妨げる要因を表 1 の左列にまとめる．文書の品質，すなわちライティングの観点において，Yan ら [10] は，プライバシーポリシーの可読性が下がる文構成や文法を整理し，半数以上のプライバシーポリシーにそのような記述が含まれていることを示した．またプライバシーポリシー特有の性質という観点で，Tang ら [4] はプライバシーポリシーに用いられる技術的用語を理解できないユーザの存在を示し，Andow ら [6] は，単一のプライバシーポリシー内でも矛盾した表現を含むことを明らかにした．原ら [11] は，収集されるデータと収集目的の対応関係を調査し，松尾ら [12] は，抽象的な記載がユーザの誤解を招くと主張した．以上に加え，我々の以前のプライバシーポリシー研究を通じた観察 [9] から，新たに 4 つの要因を追加した．

表 1 に記載した先行研究は，主にユーザスタディに基づいて要因を特定している．ユーザスタディは，ユーザの理解や認知を調査するための第一の選択肢であるが，さまざまなサービスを公開し，また随時の仕様変更を行う企業組織にとって，その都度プライバシーポリシーを最適化するためにユーザを募ることは，金銭的および時間的なコストを非常に多く負担する必要がある．

2.2 プライバシーポリシーと LLM

LLM はトランスフォーマーアーキテクチャに基づいており，Web ページや書籍などの膨大なテキストデータで事前学習された言語モデルである．従来より自然言語処理によってプライバシーポリシーの要約 [13] や，コンプライアンスをチェック [9] するための研究がなされているが，LLM のタスク遂行性能は従来技術を凌駕し，ユーザ認知の模擬 [14], [15] や法律分析 [16] においても高い適性が示されている．

表 1 プライバシーポリシーの理解を妨げる主な要因

理解を妨げる要因	参照した文献	ポリシーへの適用有無 A	B	C	D	E	各難読化が影響を与える設問 A	B	C	D	E
ライティングに起因する問題											
二重否定を用いている	Yan et al. [10]	−	✓	✓	−	−		2	2		
文または章あたりの単語数が多い	Yan et al. [10]	−	✓	−	−	−		3			
情報の提示順が論理的でない	Yan et al. [10]	−	✓	−	−	−		2			
関連情報を分散して記載している	独自	−	✓	−	✓	✓		2		1,2	2,5
プライバシーポリシー特有の性質に起因する問題											
専門用語を利用している	Tang et al. [4]	−	−	✓	−	−			3		
パラグラフ間で記載内容が矛盾している	Andow et al. [6]	−	−	−	✓	−				1,2,4	
情報の欠如（情報の対応関係無し）	原ら [11]	−	−	−	−	✓					1,4
情報の欠如（抽象的な表現の利用）	松尾ら [12]	−	✓	✓	✓	✓		7	1,2,4,7	7	4,7
前節を参照することで情報を省略している	独自	−	−	✓	−	✓			2,4		2,4
実施しない事項の記載がある	独自	−	−	✓	−	−			2,5		
既知の情報または不必要な情報の記載がある	独自	−	−	✓	✓	✓			6	6	6

LLM の急速な進歩は，プライバシーポリシーを自動分析するための新たな機会をもたらした．Tang ら [17] は，ChatGPT および GPT-4 を用いて，GDPR と関連するプライバシーポリシーの記述を高精度に分類できることを示した．Palka ら [18] は，LLM に解釈させることを前提とした網羅性重視のプライバシーポリシーのフォーマットを新しく提案し，GPT-4 によって高い精度でプライバシーの取り扱いを判別できることを示した．これらの研究は，LLM が人間の能力を超え，できる限り精緻にプライバシーポリシーを理解することを目指している．一方で，我々の目的は，プライバシーポリシーはユーザが読むために記述される文書であるという現実的な設定に即して，ユーザにとっての理解しやすさを評価し，その理解度を改善するために LLM を活用するところにある．

3. 分析対象のプライバシーポリシー

3.1 プライバシーポリシーの作成

本研究では，ユーザにとって特にセンシティブな情報 [19], [20] を取り扱うヘルスケアアプリに関するプライバシーポリシーを作成する．具体的には，位置情報や端末情報，健康状態に関するデータを取得する歩数計アプリを想定する．プライバシーポリシーの記載事項要件を満たすため，法律の専門家が執筆した専門書のサンプルをベースに拡張した [21]．具体的に取り扱うデータと利用目的については，Google Play において人気上位のヘルスケアアプリのプライバシーポリシーを参考にして記述した．ベースラインとなるプライバシーポリシー A (以降，PP-A 等と記載する) を図 1 に示す．

3.2 プライバシーポリシーの難読化

作成した PP-A に対して表 1 に示した特徴 (以降，難読化と記載) を意図的に付与する．すべての難読化を同一のプライバシーポリシー上に適用すると，可読性が著しく低下した文書となるため，表 1 の中列に示すとおり，各難読化を選択的に適用した 4 つの PP-B，C，D，E を作成した．ベースとなる PP-A には難読化を施さず，PP-B には主にライティングに関する難読化，PP-C，D，E にはそれぞれ専門用語の利用，情報の分散，情報の欠如を主としたプライバシーポリシー特有の性質に関する難読化を施した．具体的には，論理的ではない情報の提示順として，取得情報に関する記載の前に第三者への情報提供や利用者情報の送信停止に関する説明を記載した．PP-A では「外部機関による情報セキュリティマネジメントシステムに関する国際認証」と記載したが，PP-C では「ISO27001 認証」という専門用語を用いた．その結果，PP-A，B，C，D，E の文字数は，それぞれ 1,526 文字，1,502 文字，1,696 文字，1,720 文字，1,797 文字となった．

プライバシーポリシー

XXX株式会社（以下、「当社」といいます。）は、本アプリを通じて取得する利用者情報の取扱いについて、次のとおり、プライバシーポリシー（以下「本ポリシー」といいます。）を定めます。

1. 取得情報および利用目的
当社は、本アプリにおいて、以下に定める利用者情報を以下の目的のために利用いたします。なお、情報をご提供いただけない場合、本アプリおよび本サービスの全部あるいは一部をご利用いただけないことがございます。
・メールアドレス、パスワード：アカウント管理のため
・年齢、性別、体重：年齢、性別及び体重の情報から、ユーザーの歩行数に応じた消費カロリーを予測するため
・位置情報：移動距離を測定するため
・広告識別子：広告配信、広告の効果測定のため
・端末アクティビティ情報：ユーザの歩行数を推定するため
・利用者の本アプリ上での行動：本アプリの機能に対するニーズの把握のため、本アプリ上で発生しうる不具合及びその原因の特定のため、新しいサービスを開発するため

2. 取得方法
利用者による提供：メールアドレス、パスワード、年齢、性別、体重
自動取得：位置情報、広告識別子、端末アクティビティ情報、利用者の本アプリ上での行動

3. 利用者情報の提供
当社は以下の場合にのみ利用者の情報を第三者に提供いたします。
・あらかじめご本人様の同意を得た場合
・法令に基づく提供
・個人情報保護法上、本人の同意を得ることなく第三者への提供が認められる場合

ただし、当社は、取得した利用者情報を以下の範囲内で共同利用しております。
・共同して利用される個人データの項目：メールアドレス
・共同して利用する者の範囲：当社ならびに子会社および関連会社
・利用する者の利用目的：アカウント管理のため
・当該個人データの管理について責任を有する者：XXX株式会社 [住所] [代表者氏名]

4. 安全管理措置
・組織的安全管理措置：
個人情報管理責任者の設置及びその役割を明確化しています。
守秘義務の対象となる情報の流出等が発生した場合の報告体制を整備しています。
情報セキュリティ内部監査や、外部機関による情報セキュリティマネジメントシステムに関する国際認証の維持に関する審査を受けています。
・人的安全管理措置：
従業員に対して情報の秘密保持に関する誓約書の提出を義務付けています。
情報セキュリティに関する教育を継続的に実施しています。
・物理的安全管理措置：
個人情報を取り扱う区域の入退室管理を行っています。
個人情報を取り扱う機器、書類等の盗難又は紛失等を防止するための措置を講じています。
・技術的安全管理措置：
サーバーやその他情報機器においてアクセス管理を行なっています。
外部からの不正アクセスまたは不正ソフトウェアから保護する仕組みを導入しています。
システムの安全性について定期的な見直しを行なっています。

5. 利用者情報の送信停止
本アプリは、利用者情報の自動送信を停止する手段を提供しておりません。利用者情報の自動送信を停止したい場合は、本アプリをアンインストールしてください。

6. お問い合わせ窓口
ご意見、ご質問、苦情のお申出、利用者情報の開示、訂正・追記・削除、利用停止・消去のお問い合わせは、こちらのお問い合わせフォームからご連絡をお願いいたします。

7. 本ポリシーの変更
当社は、本ポリシーを、随時変更する場合があり、変更した場合には、本アプリ上に掲載します。お客様は、本アプリ上に掲載される最新のポリシーの内容を十分にご確認ください。

制定 2024.4.1

図 1 プライバシーポリシー A

表 2 設問文と選択肢

#	設問文	選択肢
Q1	XXX 株式会社によって取得、利用される情報項目とその目的として正しいと判断できる選択肢を全て選んでください。ただし、E を選ぶ場合は、他の選択肢は選ばず E のみを回答してください	A. メールアドレスをアカウント管理のために利用する B. 位置情報を広告配信のために利用する C. 端末アクティビティ情報を歩行数推定のために利用する D. メールアドレスを広告配信のために利用する E. プライバシーポリシーの内容からは判断できない（記載されていない）
Q2	XXX 株式会社から他の事業者へ提供された利用者情報の利用目的として、正しいと判断できる選択肢を全て選んでください。ただし、D を選ぶ場合は、他の選択肢は選ばず D のみを回答してください。	A. 広告の効果測定のため B. 新しいサービスを開発するため C. お客様へのご連絡のため D. プライバシーポリシーの内容からは判断できない（記載されていない）
Q3	利用者情報を安全に管理するために XXX 株式会社が行なっている措置として正しいと判断できる選択肢を全て選んでください。ただし、D を選ぶ場合は、他の選択肢は選ばず D のみを回答してください。	A. 個人情報を取り扱う機器の紛失防止のための措置 B. 外部機関によるセキュリティに関する認証の維持 C. 業務委託先の監督 D. プライバシーポリシーの内容からは判断できない（記載されていない）
Q4	XXX 株式会社と子会社間において共同利用される情報項目とその目的として正しいと判断できる選択肢を全て選んでください。ただし、D を選ぶ場合は、他の選択肢は選ばず D のみを回答してください。	A. 利用者への連絡のため、メールアドレスを利用する B. アカウント管理のため、メールアドレスを利用する C. 消費カロリー予測のため、年齢、性別を利用する D. プライバシーポリシーの内容からは判断できない（記載されていない）
Q5	プライバシーポリシーの記載内容から判断すると、業務委託は行われていますか？正しいと判断できる選択肢を 1 つ選んでください。	A. 業務委託が行われている B. 業務委託は行われない C. プライバシーポリシーの内容からは判断できない（記載されていない）
Q6	プライバシーポリシーの記載内容から判断すると、利用者情報が保管されるデータセンターの所在地はどこですか？ 正しいと判断できる選択肢を 1 つ選んでください。	A. 日本 B. 日本以外の国 C. プライバシーポリシーの内容からは判断できない（記載されていない）
Q7	プライバシーポリシーの記載内容から判断すると、データ削除の問い合わせ先はどこですか？正しいと判断できる選択肢を 1 つ選んでください。	A. XXX 株式会社の代表者 B. 問い合わせフォーム C. プライバシーポリシーの内容からは判断できない（記載されていない）

3.3 プライバシーポリシーの理解度確認アンケート

プライバシーポリシーの理解度は，記載内容に関する設問への回答によって測定する．設問文および選択肢を表 2 に示す．正解の選択肢は PP-A，B，C，D，E それぞれで異なるとともに，表 1 の右列に示したとおり，各設問に関連する記載内容に適用された難読化も異なる．なお，各設問への回答を比較することで難読化の影響を評価できるよう，難読化の割り当てを調整した．また，LLM に対して回答の根拠としたプライバシーポリシーの箇所と回答理由，参加者に対して回答の根拠としたプライバシーポリシーの箇所を尋ねた．これら根拠の箇所および回答理由をもとに，誤回答の傾向を分析する．

4. LLM による理解度評価

RQ1: LLM はどの程度プライバシーポリシーを理解できるか？を明らかにするため，複数の LLM に対して分析対象のプライバシーポリシーと設問をプロンプトとして入力し，LLM による出力を評価する．

4.1 使用モデル

2024 年 7 月時点で OpenAI，Anthropic，Google が公表する最新モデル gpt-4o-2024-05-13，claude-3-5-sonnet-20240620，gemini-1.5-pro を API 経由で利用する．LLM の出力の再現性を確保するため，ランダム性を調整するパラメータ Temperature を 0.0 に設定し，出力の分散を抑えた．なお，GPT 以外の現在の LLM では．シード値を固定できないため，本実験では採用しなかった．

4.2 プロンプトの設計

後述するユーザスタディと条件を揃えるため，事前知識を与えない Zero-Shot プロンプティングを採用する．LLM への入力は「以下は，架空のスマホの歩数計アプリのプライバシーポリシーです.」という説明文，プライバシーポリシー本文，設問文，選択肢，出力形式を指定する文言から構成される．1 回の入力には，3 章で作成したプライバシーポリシーのうち 1 種類，表 2 の設問のうち 1 問が含まれる．出力は LLM の回答，回答の根拠としたプライバシーポリシーの箇所，回答の理由の説明から構成される．出力形式は JSON 形式に統一するため，「次の JSON 形式で回答してください。回答 例:`{"Answer":["A","B","C","D","E"],"Citation":string,"Reason":string}`」とプロンプトで指定した．なお，同じ設問を 10 回繰り返し回答させ，各設問への正答率によって理解度を測定する．

4.3 結果

4.3.1 LLM の理解度

モデルごとの各設問への正答率を表 3 に示す．すべてのモデルにおいて，難読化を含まない PP-A およびライティングに関する難読化を含む PP-B への正答率は 100%であり，LLM はライティングに関する難読化の影響を受けな

表 3 LLM による各設問への正答率（%）

PP	モデル	Q1	Q2	Q3	Q4	Q5	Q6	Q7	合計
A	GPT	100	100	100	100	100	100	100	100
	Gemini	100	100	100	100	100	100	100	100
	Claude	100	100	100	100	100	100	100	100
B	GPT	100	100	100	100	100	100	100	100
	Gemini	100	100	100	100	100	100	100	100
	Claude	100	100	100	100	100	100	100	100
C	GPT	0.0	80	100	0.0	100	100	100	69
	Gemini	0.0	0.0	100	0.0	100	100	0.0	43
	Claude	0.0	0.0	100	0.0	100	100	100	57
D	GPT	0.0	100	100	100	100	100	100	86
	Gemini	0.0	100	100	90	100	100	0.0	70
	Claude	100	100	100	0.0	100	100	100	86
E	GPT	0.0	100	100	0.0	100	100	100	71
	Gemini	0.0	100	100	10	100	100	0.0	59
	Claude	0.0	0.0	100	0.0	100	100	100	57

かった．一方，プライバシーポリシー特有の難読化を含む PP-C，D，E への正答率は低かった．また，3 件の例外を除き各設問の正答率は 0%または 100%であり，LLM の回答は概ね安定していることが示された．

4.3.2 LLM の誤回答の傾向

情報の分散の影響．PP-D の難読化では，データの取り扱いに関する記述の一部をファーストパーティではなく共同利用の章で記載した．したがって，Q1 のデータ取り扱いに関する設問では，取得情報および利用目的に関する章に記載されている内容と共同利用に関する章に記載されている内容を合わせて回答する必要があり，選択肢 A，C，D をすべて選ぶことが正解である．Claude は 2 つの章を特定して，3 件の選択肢をすべて回答した．一方，GPT と Gemini はファーストパーティによる取得情報および利用目的の章のみを引用し，選択肢 A と C のみを回答した．複数の情報を踏まえて回答する必要があったとしても，最も関連のある情報のみを参照し，他の情報は参照しないことがわかる．

情報の欠如（対応関係なし）の影響．PP-E の難読化では，取得する情報と利用目的を分けて記載し，どの情報項目がどのような目的で利用されるかの対応関係を明示しなかった．したがって，取得する情報項目とその利用目的の組み合わせを問う Q1 の正解は，選択肢 E「判断できない（記載されていない）」である．しかし，GPT と Claude は「[選択肢の情報項目] は取得情報として明記されており，利用目的に [選択肢の利用目的] が含まれている」ことを理由に誤答した．対応関係が明示されていない場合に，LLM は「判断できない（記載されていない）」と回答せず，取得情報の章に記載された情報はすべて利用目的の章に記載されている目的で使用されると誤解することがわかった．

情報の欠如（抽象表現）の影響．PP-C の難読化は，利用者情報の利用目的に関する記述を「サービスの基本機能のため」という抽象的な表現に変換している．本研究では，歩数計アプリの「基本機能」を「歩行数推定」および「消費カロリー予測」であると定義し，利用目的に関連する PP-C の Q1 と Q4 (以降，PP-C-Q1 や PP-C-Q4 等と記載する) について，それぞれ選択肢 A，C および選択肢 B，C を正解とした．しかし，GPT（Q1，Q4），Claude（Q1，Q4），Gemini（Q1）は，抽象的な表現を解釈することなく，「明確な記載はない」と回答していた．LLM は情報の欠如を検知することができ，ユーザの理解を阻害する可能性があるプライバシーポリシーの箇所を特定できると考えられる．

PP-B，C，D，E の難読化では，問い合わせ先の情報に関して「その他」という抽象的な表現を用いて，「ご意見，ご質問，苦情のお申出その他利用者情報の取扱いに関するお問い合わせは，こちらのお問い合わせフォームからご連絡をお願いいたします。」と記載した．「その他利用者情報の取扱い」にデータの削除が含まれる前提のもと，その問い合わせ先を尋ねる Q7 の正解は選択肢 B とした．Gemini は，PP-C，D，E では「その他」の範囲を解釈せず，「判断できない（記載されていない）」と回答したが，PP-B では正解である選択肢 B を選んだ．抽象表現による難読化の他に，PP-B はライティングに起因する難読化，PP-C，D，E はプライバシーポリシー特有の性質に起因する難読化を主に適用しているため，抽象表現の理解は前後の文脈に影響を受けることがわかった．また，GPT と Claude は「サービスの基本機能のため」という抽象表現を解釈しなかったが，「その他利用者情報の取扱い」という抽象表現は正しく解釈した．LLM による抽象表現の理解は，抽象化の度合いによって異なる可能性が示された．

参照による情報の省略の影響．PP-C，E の難読化では，共同利用によるデータ取り扱いについて，前の章を参照することにより記載を省略した．他の事業者への情報提供に関する Q2，Q4 では，参照元であるファーストパーティによる取得情報および利用目的の章に記載されているデータ取り扱い方法を回答することが正解である．しかし，GPT（PP-C-Q2）は 10 回の試行中 1 回不正解であり，参照元に記載されている情報の一部を回答したが必要な情報をすべて選択することはできなかった．また，Gemini（PP-C-Q2）は参照元を確認せず，選択肢 A，B，C のいずれも正しいと判断しなかった．一方，PP-E-Q2 は GPT，Gemini ともに正解していた．PP-C，E の参照元の記載の大きな違いは，PP-C では設問に関係する情報と関係しない情報が混在しているのに対し，PP-E では設問に関連する情報のみが記載されていることである．プライバシーポリシーにおいて前章を参照する形式が採用されていた場合，参照元の記載方法が LLM の理解度に影響することがわかった．

専門用語の解釈の違い．Claude と GPT，Gemini 間において一部の単語の解釈が異なる事例を確認した．具体的には，Q2 の設問における「他の事業者への提供」という表

現について，GPT と Gemini は第三者提供，共同利用，業務委託等あらゆる形態の他の事業者との連携を考慮して回答したが，Claude は「共同利用の範囲についても言及されていますが，これは他の事業者への提供ではなく（以下省略）」と回答した．また，プライバシーポリシー上で個人情報の利用目的の 1 つとして記載した「アカウント管理」という表現について，GPT と Gemini はアカウント管理には利用者への連絡は含まれないと解釈していたが，Claude だけが「アカウント管理には利用者への連絡も含まれる」と解釈をして，誤った選択肢を回答した事例があった．

5. ユーザによる理解度評価

RQ2: ユーザはどの程度プライバシーポリシーを理解できるか？を明らかにするため，クラウドソーシングサービスで募集した参加者に対して，分析対象のプライバシーポリシーと設問を用いたユーザスタディを実施する．

5.1 参加者募集

Lancers を用いて参加者を合計 500 名募集した．募集は 2024 年 6 月，7 月に実施し，幅広い属性の参加者を募るために平日と休日に分けて募集を行った．アンケート回答の所要時間を 20 分と想定[*1]し，時給換算で日本の全地域の最低賃金を超える 400 円を報酬額とした．

5.2 アンケートの設計

アンケートの内容は，参加同意，架空アプリのプライバシーポリシーの提示，プライバシーポリシーの理解度を問う質問，参加者の属性を問う質問の 4 つの大項目から構成される．参加者は 3 章で作成したプライバシーポリシーのうち 1 種類を読み，表 2 に示した 7 種類の設問に回答する．参加者の属性を問う質問として，年齢，性別などのデモグラフィックスおよびプライバシーに関する経験について尋ねた．なお，回答の質を担保するために，注意力を確認する簡単な質問を設け，誤答した場合は無効回答とした．

5.3 結果

5.3.1 参加者の属性

PP-A，B，C，D，E に対する各 100 件の回答を募集し，無効回答を除外した結果，それぞれ 88 件，94 件，82 件，92 件，93 件の回答が得られた．参加者の属性およびプライバシーに関する経験の統計を表 4 に示す[*2]．すべてのプライバシーポリシーにおいて，70%以上の参加者が PC からアンケートに参加していた．参加者の専門性は，「行政・法学」が 1–4%と低かったが，「情報・通信」は 8–18%と高かった．プライバシーポリシーを読むタイミングは「サービス登録時」，その読み方はキーワード検索やツール等はあまり使わず 「読み流す」参加者が最多であった．

表 4 参加者の属性およびプライバシー関連の経験に関する統計 (%)

	A	B	C	D	E
利用デバイス					
PC	72.7	74.5	74.4	68.5	79.6
スマートフォン	22.7	24.5	23.2	23.9	18.3
タブレット	4.5	1.1	1.2	7.6	2.2
その他	0.0	0.0	1.2	0.0	0.0
専門領域					
情報・通信	18.2	11.7	15.9	7.6	11.8
行政・法学	3.4	1.1	2.4	2.2	4.3
その他	68.2	84.0	76.8	83.7	72.0
無回答	10.2	3.2	4.9	6.5	11.8
プライバシーポリシーを読むタイミング					
サービス登録時	79.5	72.3	70.7	78.3	75.3
更新通知受領時	34.1	24.5	28.0	22.8	22.6
個人情報入力時	29.5	22.3	28.0	22.8	30.1
読まない	17.0	21.3	20.7	12.0	24.7
その他	0.0	1.1	1.2	1.1	1.1
プライバシーポリシーの読み方					
徹底的に読む	8.0	3.2	3.7	5.4	12.9
読み流す	67.0	72.3	69.5	68.5	69.9
キーワードを検索する	15.9	7.4	11.0	8.7	10.8
題名を参照する	26.1	19.1	17.1	18.5	18.3
ツールを利用する	1.1	0.0	2.4	1.1	0.0
読まない	17.0	17.0	18.3	14.1	18.3
その他	2.3	5.3	0.0	2.2	0.0

表 5 ユーザによる各設問への正答率 (%)

PP	Q1	Q2	Q3	Q4	Q5	Q6	Q7	合計
A	64.8	62.5	54.5	73.9	65.9	89.8	93.2	72.1
B	71.3	69.1	69.1	80.9	53.2	90.4	84.0	74.0
C	25.6	35.4	43.9	17.1	61.0	89.0	81.7	50.5
D	15.2	57.6	64.1	66.3	48.9	89.1	72.8	59.2
E	6.5	50.5	73.1	21.5	90.3	92.5	78.5	59.0

5.3.2 参加者の理解度

参加者の各設問への正答率を表 5 に示す．なお，「D（E）を選ぶ場合は，他の選択肢は選ばず D（E）のみを回答してください．」という指示に従っていない回答 4 件（PP-B-Q2,4 と PP-D-Q1,2 の各 1 件）は分析対象外とした．参加者による正答率は最小 6%，最大 93%，平均 63%であった．難読化を含まない PP-A およびライティングに関する難読化を加えた PP-B への正答率は 7 割を超えた．一方，プライバシーポリシー特有の難読化を含む PP-C，D，E への正答率はいずれも 6 割を下回った．

5.3.3 参加者の誤回答の傾向

情報の分散の影響．分散した情報を合わせて回答する必要がある設問（PP-D-Q1）に対する正答率は 15%と低かった．選択肢ごとの正答率としては，選択肢 A，C は設問に関連の深い章に情報が記載されていたため，その正答率は 97%，90%と高かったが，選択肢 D は関連の浅い章に情報が分散したため，その正答率は 23%と低下した．選択肢 D

[*1] 著者らによるパイロットスタディでは，回答時間は 15 分程度であった．
[*2] 紙面の都合上，分析結果に影響が見られなかったデモグラフィックスは割愛する．

を選択できなかった参加者はいずれも，関連の浅い章を参照していなかったことが回答根拠からも判明した．参加者は参照すべき情報が分散していると，関連の深い章のみを参照する傾向があった．

情報の欠如（対応関係なし）の影響．ファーストパーティによって取得する情報項目と利用目的の対応関係がないPP-E を読んだ参加者に対して，情報項目とその利用目的を質問した結果，正答率は 6.5%と低かった．6.5%の参加者は，関連する章を参照しながら「取得情報を何のために利用するか明記されていないから」,「2. 取得情報 と 3. 利用目的 の内容から，どの情報をどの目的に利用するという具体的な結び付けの情報は明示されていないため.」とその根拠を回答したが，それ以外の参加者は情報項目と利用目的の対応関係について各自の解釈を加えて回答した．必要な情報が欠如している場合，ユーザが各自の解釈を加えることにより，企業とユーザの間で認識の齟齬が発生する可能性があることが示された．

情報の欠如（抽象表現）の影響．抽象的な表現を使った記述に関する設問 PP-C-Q1 および PP-C-Q4 における参加者の正答率はそれぞれ 26%，17%であった．Q1 および Q4 どちらも，抽象表現が関連しない選択肢の正答率はそれぞれ 96%と 76%と高かった一方で，関連する選択肢の正答率は 52%と 30%と大きく下回った．

参照による情報の省略の影響．PP-C，E の Q2 は前節の参照によって省略されている情報に関する設問である．PP-C-Q2 と PP-E-Q2 どちらも正解の選択肢は，プライバシーポリシー内に同一の文言が記載されているため，検索機能を用いると容易に正解を導出できるが，完答した参加者はそれぞれ 35%，51%であった．この結果は，表 4 に示したとおり，参加者の 7 割以上が PC を利用していた一方で，キーワード検索を活用してプライバシーポリシーを読む参加者は 10%前後しかいなかったという結果とも整合が取れる．また，不正解だった参加者の回答根拠を確認すると，約 7 割が前節を参照できていなかったため，参照による省略によってユーザの理解度が下がることがわかった．

専門用語の知識不足．専門用語を使って記載した PP-C は，専門用語を使っていない PP-A，B，D，E と比較して Q3 の正答率が低かった．プライバシーポリシーにおける専門用語の使用がユーザの理解度を下げる要因になることがわかる．また，PP-C-Q3 の正答/誤答した参加者数と「情報・通信」または「行政・法学」/それ以外の専門性の参加者数の間でカイ二乗検定を行なった結果，有意差（$p < 0.05$）が確認され,「情報・通信」または「行政・法学」の専門性を有する参加者は，専門用語を理解して正答していたことがわかった．

PP-B，D ではプライバシーポリシー上で業務委託の有無について言及していないが,「業務委託が行われている」や「行われていない」という誤回答は，PP-B において 30%，17%，PP-D において 39%，12% それぞれ確認された．前者の回答根拠のうち，86%が共同利用，14%が第三者提供，6%が安全管理措置に関する記述を参照していた．後者の回答根拠も，56%が共同利用，15%が第三者提供，30%が安全管理措置に関する記述を参照していた．業務委託と共同利用を同一視している参加者が多いことがわかった．

表 6 LLM（GPT）とユーザの理解度比較．○: 両者とも理解，◐: LLM のみ理解，●: 両者とも理解不足．

PP	Q1	Q2	Q3	Q4	Q5	Q6	Q7
A	○	○	○	○	○	○	○
B	○	○	○	○	○	○	○
C	●*1	◐*2	◐*3	●*2	○	○	○
D	●*2	○	○	○	◐*3	○	○
E	●*4	○	○	●*4	○	○	○

*1: 情報の欠如 (抽象表現)，*2: 情報分散
*3: 専門用語，*4: 情報の欠如（対応関係なし）

情報の見逃し．難読化が適用されていない PP-A の Q3 への正答率は 55%と低かった．PP-A に記載されている安全管理措置の内容は，正解の選択肢 A，B とほぼ同一の文言を含んでいた．誤答した参加者のうち 88%は安全管理措置に関する章を適切に参照していたが，不注意で情報を見逃してしまったと考えられる．

6. 議論

6.1 理解度測定の自動化に向けて

4 章と 5 章の結果を対比することで，LLM とユーザにおける誤回答の共通点と相違点が明らかとなる．LLM とユーザの理解度を比較した結果を表 6 に示す．正答率が 50%を超える設問を「理解」と分類した．ユーザより LLM のほうが総じて正答率が高かった．文書ごとにみると，ユーザと LLM が理解できたもの (○)，両者が理解できなかったもの (●)，LLM のみが理解できたもの (◐) が存在する．すなわち，PP-A,B は読解上の問題がなく，PP-E では LLM でユーザ認識の齟齬を特定でき，PP-C,D では LLM とユーザの間に理解のギャップが生じている (LLM でユーザ認識の齟齬を特定できない) ことを意味する．

情報の分散および情報の欠如による難読化は，LLM とユーザにとって同様に理解を妨げることがわかった．こうした欠陥を含んだプライバシーポリシーにおいて，LLM が理解できない箇所を自動的に特定することで改善すべき記述を洗い出し，ユーザにとっての理解度を高めることに寄与できる．一方，専門用語の知識不足や情報の見逃しによる誤回答は，LLM では観測されずユーザにおいてのみ観測された．こうしたプライバシーポリシーにおいては，Zero-Shot プロンプティングによって指示された LLM では欠陥を認識できないため，他の手法で補完する必要がある．専門用語の使用によりユーザが理解しにくい記述を特定するための手法として，先行研究 [7] で明らかにされた

専門用語に対するユーザの理解度や誤解の内容を踏まえ，専門用語の分類器を作成することができる．また，LLM はペルソナを付与することで回答タスクの精度や傾向を調整できることが知られており [22]，専門用語を理解しないという前提で回答を出力させるアプローチが考えられる．

6.2 制約と課題

本研究ではプライバシーポリシーに対して複数の難読化を含め，LLM とユーザの理解度を測定し，理解できなかった要因を特定した．ただし，難読化ごとの影響度の大きさや，難読化の組み合わせによる影響の変化については，調査対象としなかった．また，4.3.2 項では，同一の難読化であっても，前後の文章によって LLM が受ける影響が異なる可能性が示唆された．LLM が難読化の影響を受ける条件をより明確にすることで，再現性を持ってユーザが理解しにくい箇所を特定できると考える．

ユーザスタディでは「生成 AI の利用を禁止する」ことを指定してタスクを依頼したが，参加者の回答の中には LLM の回答と類似するものが含まれていた．このような LLM を補助的に使用した回答の排除は，注意力確認テストによる対策では困難なため，LLM による回答を検知する新しい確認テストを導入する必要があると考える．

6.3 研究倫理

本研究のユーザスタディは架空のプライバシーポリシーへの理解度を評価するためのものであり，個人情報などの機微な情報は取得していない．加えて，アンケートのはじめに参加同意を取得した．具体的には，アンケートへの回答はいつでも辞退できること，アンケート結果は匿名データとして扱われ，研究目的でのみ使用されることを明記し，同意を得た．なお，ユーザスタディの内容は我々の組織内の倫理委員会によって審査され，承認を受けている．

7. おわりに

本研究では，11 種類の難読化を含む複数のプライバシーポリシーを準備し，それらに対して 3 種類の LLM および 449 名のユーザによる理解度を比較分析した．その結果，LLM の理解度はユーザのそれを総じて上回るとともに，LLM およびユーザどちらの理解度も下げる要因 (情報の分散や情報の欠如など) を特定した．加えて，専門用語の知識不足や情報の見逃しなど，ユーザ特有の誤解を招く要因があることも明らかにした．ユーザのみが理解できない記述が存在するものの，LLM が誤解する記述はユーザも誤解するため，LLM を用いることによりプライバシーポリシーの理解度測定を自動化できる可能性を示した．今後は LLM とユーザ間の理解度のギャップをさらに埋めるために，ユーザの誤り傾向について理解を深めるとともに，LLM に対して前提条件やペルソナ等を付与することにより，ユーザの考えや振る舞いを模擬する方法を検討したい．

参考文献

[1] Data Protection Commission, “Data Protection Commission Announces Decision in WhatsApp Inquiry,” 2021.

[2] Bloomberg News, “Amazon Given Record 888 Million EU Fine for Data Privacy Breach,” 2021.

[3] A. Micheti *et al.*, “Fixing Broken Doors: Strategies for Drafting Privacy Policies Young People Can Understand,” *Bulletin of Science, Technology & Society*, vol. 30, no. 2, pp. 130–143, 2010.

[4] J. Tang *et al.*, “Defining privacy: How users interpret technical terms in privacy policies,” in *PoPETs*, 2021.

[5] K. Mori *et al.*, “Analysis of Privacy Compliance by Classifying Policies Before and After the Japanese Law Revision,” *Journal of Information Processing*, vol. 31, pp. 829–841, 2023.

[6] B. Andow *et al.*, “PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play,” in *USENIX Security Symposium*, 2019.

[7] 金森ら, “プライバシーポリシーに使用される技術用語および個人情報保護法に対するユーザの理解度の調査,” in **コンピュータセキュリティシンポジウム**, 2023.

[8] C. Jensen *et al.*, “Privacy practices of Internet users: Self-reports versus observed behavior,” *International Journal of Human-Computer Studies*, vol. 63, no. 1-2, pp. 203–227, 2005.

[9] K. Mori *et al.*, “Analysis of Privacy Compliance by Classifying Multiple Policies on the Web,” in *COMPSAC*, 2022.

[10] C. Yan *et al.*, “On the quality of privacy policy documents of virtual personal assistant applications,” in *PoPETs*, 2024.

[11] 原ら, “プライバシーポリシーにおける収集データと目的の対応関係の実態調査,” in *SPT* **研究会**, 2024.

[12] 松尾ら, “利用規約・プライバシーポリシーの作成・解釈ーー国内取引・国際取引を踏まえて.” 商事法務, 2023.

[13] R. N. Zaeem *et al.*, “Privacycheck: Automatic summarization of privacy policies using data mining,” *ACM TOIT*, vol. 18, no. 4, pp. 1–18, 2018.

[14] L. Salewski *et al.*, “In-context impersonation reveals large language models’ strengths and biases,” in *NeurIPS*, 2024.

[15] Y. Shao *et al.*, “Character-llm: A trainable agent for role-playing,” *arXiv:2310.10158*, 2023.

[16] F. Yu *et al.*, “Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks,” in *ACL*, 2023.

[17] C. Tang *et al.*, “PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models,” *arXiv:2309.10238*, 2023.

[18] P. Pałka *et al.*, “No More Trade-Offs. GPT and Fully Informative Privacy Policies,” *arXiv:2402.00013*, 2023.

[19] T. Cory *et al.*, “A qualitative analysis framework for mhealth privacy practices,” *arXiv:2405.17971*, 2024.

[20] M. Rizwan *et al.*, “Risk monitoring strategy for confidentiality of healthcare information,” *Computers and Electrical Engineering*, vol. 100, p. 107833, 2022.

[21] 白石ら, “プライバシーポリシー作成のポイント.” 中央経済社, 2022.

[22] C. Olea *et al.*, “Evaluating Persona Prompting for Question Answering Tasks,” in *AIS*, 2024.