



Make privacy policies longer and appoint LLM readers

Przemysław Pałka¹ · Francesca Lagioia^{2,3} · Rūta Liepina³ · Marco Lippi⁴ · Giovanni Sartor^{2,3}

Accepted: 1 February 2025
© The Author(s) 2025

Abstract

In a world of human-only readers, a trade-off persists between comprehensiveness and comprehensibility: only privacy policies too long to be humanly readable can precisely describe the intended data processing. We argue that this trade-off no longer exists where LLMs are able to extract tailored information from clearly-drafted fully-comprehensive privacy policies. To substantiate this claim, we provide a methodology for drafting comprehensive non-ambiguous privacy policies and for querying them using LLMs prompts. Our methodology is tested with an experiment aimed at determining to what extent GPT-4 and Llama2 are able to answer questions regarding the content of privacy policies designed in the format we propose. We further support this claim by analyzing real privacy policies in the chosen market sectors through two experiments (one with legal experts, and another by using LLMs). Based on the success of our experiments, we submit that data protection law should change: it must require controllers to provide clearly drafted, fully comprehensive privacy policies from which data subjects and other actors can extract the needed information, with the help of LLMs.

Keywords Privacy policies · Artificial intelligence · Large language models · GDPR compliance · Question answering · Information retrieval

1 Introduction

Imagine you just listened to a song or ordered food using an app. Now, picture asking your smartphone: “Could you tell me the names of all the corporations who will know that I just listened to *Frankly, Mr. Shankly* by The Smiths, and ordered baked tofu, and list all the purposes for which they will use this data?”¹ As a consumer, you have the right to know this (Calo 2011; Kaminski 2019; Jabłonowska and Tagiuri 2023). Such knowledge is a precondition for choosing whether and how to use

¹ These are no (longer) trivial matters; in a ML-powered inference economy (Solow-Niederman 2022) such data can be proxy for estimating one’s race (Marshall and Naumann 2018) or political convictions (DellaPosta et al 2015).

Extended author information available on the last page of the article

various services (Reidenberg et al 2015). And yet, in 2025, receiving a straightforward answer to such a question is not possible.

This is due to the fact that current privacy policies *do not contain* the information which is necessary to answer such straightforward questions. Our empirical research reveals that privacy policies of online services are full of vague terms (“we collect data about your use of our service”), open-ended catalogs (“such as,” “including,” “for example”) and do not specify what categories of data are shared with whom (“we share your personal data with our marketing partners”).² Consequently, even a hypothetical human expert with unlimited resources would not be able to answer simple questions, like those from the first paragraph, based on the content of existing privacy policies.

The law³ is (partly) to blame for this state of affairs. It requires privacy policies to simultaneously satisfy two requirements: to be *comprehensive*, i.e., to contain information about all the data categories processed, purposes of use, and their recipients; and to be *comprehensible*, i.e., to be concise and written in simple non-technical language. In a world of human-only readers, there is a trade-off between the two requirements: more comprehensible policies tend to be less comprehensive, and vice versa. The comprehensibility requirement has been seen as the more important one, and advocated for in scholarship, calling for ever more simplification (Calo 2011; Waldman 2018; Emami-Naeini et al 2021; Yerby and Vaughn 2022).

In this paper, we challenge the conventional wisdom and adopt the opposite approach. We argue that the law should focus on requiring privacy policies to be fully comprehensive, *even if* at the cost of making these documents much more extended. The reason is simple: soon, anyone could engage in simplification of, or information extraction from, long privacy policies, as such efforts can be increasingly automated through the use of large language models (LLMs). Of course, LLMs are not sufficient; we don’t expect the majority of consumers to now be copy-pasting privacy policies into ChatGPT and conversing about them with a machine. Clearly, additional technological solutions, like automated warnings or specialized chatbots could support consumers. Moreover, LLMs and other tools could be relevant not only to consumers but also to collective bodies (such as consumer associations) interested in evaluating privacy policies and data processing practices (Lippi et al 2020). However, for all these tools to provide answers technological performance is not sufficient; it is also required that privacy policies contain comprehensive information about corporate practices concerning consumers’ personal data.

Thus, our work is meant to test the following (counterfactual) hypothesis: if privacy policies were fully comprehensive, then LLMs could provide adequate answers to consumer queries concerning the processing of personal data by corporate actors.

To verify this hypothesis we proceeded as follows. First, we wrote a mock privacy policy of “Orderoo Inc.,” a company similar to DoorDash, Deliveroo, etc., in the format we propose as fully comprehensive. This format and the methodology adopted are also novel contributions presented in this paper. Their logic is discussed

² See *infra* Sect. 5.

³ In this paper we focus on the EU law, i.e. the GDPR, while also accounting for the “notice and choice” model prevalent in the U.S. See *infra* Sect. 3.

below,⁴ an excerpt presented in Appendix A, and the full privacy policy, with all the experiment documentation, made available on GitHub.⁵ Then, we came up with a list of 9 specific questions a consumer might be interested in knowing the answers to. Finally, we tested the extent to which two LLMs, i.e., GPT-4 and Llama2-7B, can answer these consumer questions based on the fully comprehensive Orderoo-Inc privacy policy.⁶

The results are promising. Even though Llama's performance has proven far from satisfactory, GPT-4 has been able to answer six out of nine questions flawlessly, while keeping F1 score above 85% for the three remaining questions as regards the best performing prompts.

We also asked the LLMs the same 9 questions based on the (non-comprehensive and ambiguous) privacy policies currently used by 5 food-delivery companies, i.e., Deliveroo, DoorDash, Glovo, Just Eat and Wolt. The results were significantly worse, with GPT-4 answering 9 questions correctly out of forty-five, and Llama2 giving only one correct answer. This demonstrates that even though GPT-4 is already capable of answering consumers' questions based on a well-structured and fully comprehensive privacy policy, using it to answer questions based on the existing policies might not only be ineffective but also dangerous (as it might mislead consumers). This further strengthens our policy suggestion and underscores the need to move towards fully comprehensive privacy policies as soon as possible.

The paper is divided into nine sections. First, we discuss the related work, and second, the legal background. Third, we describe the test case, i.e., the questions for which a consumer might reasonably expect to find answers when consulting a privacy policy. Fourth, we demonstrate that the existing privacy policies do not contain the necessary information. Fifth, we discuss what a fully comprehensive privacy policy could look like. Sixth, we report the results of our experiments. Seventh, we show how the LLMs fare when asked the same questions based on the actual privacy policies. Finally, we discuss implications for further research in computer science, law, and policy.

2 Related work

The automated analysis of privacy policies necessitates an interdisciplinary approach that integrates law and AI. In this section, we highlight the work done at this intersection and underscore potential gaps. We focus on three lines of relevant research: empirical legal research on privacy policies; AI & law research on NLP in the legal domain; and recent attempts to apply LLMs to extract information from legal documents. Additionally we consider also some contributions illustrating and comparing EU and U.S. law on privacy and data protection.

⁴ See *infra* Sect. 6.

⁵ <https://github.com/marcolippi83/comprehensive-privacy-policies>.

⁶ See *infra* Sect. 7.

2.1 Empirical research into privacy policies

Empirical legal research has conclusively argued that privacy policies are too often ineffective: on the one hand, they are not read by data subjects since they are too long and complex; on the other hand, they do not provide data subjects with information sufficiently precise to be actionable. It has indeed been shown that consumers do not read privacy policies (Bakos et al 2014; Obar and Oeldorf-Hirsch 2020) and, when they do, they often misunderstand them (Reidenberg et al 2015). It has been estimated that reading the privacy policies of all websites one uses would take hundreds of hours each year (McDonald and Cranor 2008). Researchers have argued that privacy policies are too long and too confusing for individuals to understand them (Yerby and Vaughn 2022; Wagner and Walker 2019). As a solution to this problem, numerous scholars have suggested making privacy policies shorter or easier to understand, e.g., by using the so-called “privacy nutrition labels” or relying on other means of simplification (Calo 2011; Emami-Naeini et al 2021; Kelley et al 2009; Waldman 2018). However, it has also been shown that even simplified privacy policies have a close-to-zero effect on consumers’ understanding of corporate data practices affecting their rights (Ben-Shahar and Chilton 2016). This last finding might suggest that the problem with privacy policies is not (only) their form but also their content. Put simply: privacy policies often *do not contain* the information necessary for consumers—even the hypothetical ones with time and expertise to read them—to learn how their data will be processed (Contissa et al 2018). A recent study (Pałka et al 2023) highlighted how, by using vague terms for data categories (“we collect data about your *use of our service*”) or purposes of processing (“we use your data to *improve our service*”), many privacy policies leave corporations the freedom to engage in data practices that consumers cannot reasonably expect or predict. The same study demonstrated that many privacy policies do not explain what data will be used for what purpose or shared with what recipient (Pałka et al 2023). In this sense, not reading privacy policies (as they are today) might very much be the rational thing to do, as there’s nothing interesting to be learned from them.

All these observations have led numerous scholars to theorize the so-called privacy paradox (Kokolakis 2017; Gerber et al 2018). This notion refers to a situation where consumers’ stated preferences suggest that they value privacy highly, but their revealed preferences indicate that they do not. Put simply: when asked in a survey, consumers claim to care about privacy but when accessing a website, they click they agree to everything, without reading anything. The idea of the privacy paradox has been criticized by some scholars (Solove 2021; Waldman 2020), arguing that it is the mix of cognitive biases, bad design, and other purposefully created obstacles that render privacy self-management too difficult. In our view, many individuals *do* care about privacy, but judge the cost of self-managing it to be higher than the expected losses from privacy violations (the severity of such losses times the probability of suffering them). We believe this is the case due to the stated preferences in the empirical studies (Kokolakis 2017; Gerber et al 2018), occasional public uproar following data breaches (Kvedar 2019; Kenny 2018), and the market practice of corporations using data for profit, but generally holding it secure, so that unwanted disclosure happens way less often than detrimental use.

If privacy is important to us, but the costs of engaging in its protection are too high, then by reducing these costs we may encourage individual and collective bodies to take action. The cost-reduction measure we are considering concerns acquiring knowledge of the processing practices of data controllers. As noted above, this measure has two prongs: requiring companies to draft fully comprehensive privacy policies and enabling users to query such policies through LLMs.

We are mindful of the fact that many other advances would be needed to ensure that consumers' privacy is effectively protected (like interfaces for automated communication of privacy preferences, or legal requirements for corporations to respect them). However, we argue that our proposal could contribute to a broader change of the status quo. For the suggested use of LLMs to have a beneficial—though limited—impact, it is not necessary that the majority of consumers start spending time querying LLMs about the contents of privacy policies. It might be sufficient that some consumers and collective bodies achieve better awareness of unwanted practices by particular controllers, so that such practices are brought to the light and duly challenged.

2.2 AI & law research on NLP in the legal domain

In our view, the effectiveness of privacy policies as communication instruments strongly depends on who (or what) could be their readers and addressees. The intended audience can indeed change with the advances in natural language processing (NLP). Several studies have shown that NLP tools can be used to analyze massive amounts of information, made available to consumers, and assist them in understanding and assessing their content. Relevant examples include summary generators (Zaeem et al 2018), document completeness checkers (Costante et al 2012) and ontologies (Palmirani et al 2018; Oltramari et al 2018). In consumer law, significant progress has been made towards automating the identification of potentially unfair clauses in online contracts and privacy policies, as part of the CLAUDETTE project (Lippi et al 2019). A similar methodology was applied to analysing privacy policies based on the General Data Protection Regulation (GDPR)⁷ (Contissa et al 2018), with less convincing results due to the complexity, incompleteness and vagueness of these documents (Liepina et al 2019). Further attempts have been made to summarize privacy policies (Zaeem et al 2018), answer users' questions (Ravichander et al 2019), or represent the content of policies in a user-friendly form (Sadeh et al. 2013; Harkous et al 2018).

However, the usefulness and the practical value of such NLP analysis—automated representation, assessment, summarization, and question-answering—could be only as good and specific as the information *actually included* in the analyzed documents. As of today, the value of this information is limited since the requirement for such policies to be simultaneously exhaustive and understandable prevents complete and precise information being provided.

⁷ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), OJ L 119, 4.5.2016, p. 1–88, hereinafter “the GDPR.”

LLMs have dramatically changed natural language processing, also extending their impact across the whole field of artificial intelligence (Chang et al 2023). These recent breakthroughs have created new opportunities for analyzing texts through simple prompt-based queries. LLMs use transformer-based architectures, and have been trained on vast amounts of general data such as webpages, books, articles, covering a wide range of topics (Vaswani et al. 2017). It has been argued that the performance of LLMs in the legal domain can be improved with fine-tuning, given that legal language is highly specialized (Chalkidis et al 2021). While these developments are fairly recent, some studies testing the abilities of LLMs in legal analytics have already emerged (Savelka et al 2023; Baron et al 2023). It has been shown that GPT models (3, 3.5, and 4) can answer the US (Katz et al 2023) and Japanese (Yu et al 2023) bar exam questions on par with the examinees. These tests include various reasoning tasks such as rule application, legal opinion drafting, and multiple-choice questions.

2.3 NLP and privacy policies

Within the growing research on NLP in the legal domain, some contributions have specifically addressed privacy policies. As for privacy policy analysis, Tang et al. (2023) experimented with ChatGPT and GPT-4 to improve the extraction from privacy policies of information, relevant relevant to the GDPR—e.g., “how and why a service provider collects user information,” “how long user information is stored?” The reported results of their PolicyGPT show a significant increase in retrieval metrics as compared to older models, which used expert annotated datasets.

Regarding the format of the privacy policies, the data processing information can be represented in several different ways. On the one hand, there are various machine-readable formats (Kumaraguru et al 2007; Tøndel and Nyre 2012) (though scholarly attempts at constructing those seem to have culminated about a decade ago), potentially based on ontologies (Gharib et al 2020; Palmirani et al 2018). On the other hand, there are human-readable formats, i.e., texts or tables. However, the advances in NLP (Zhong et al 2020) contribute to blurring the machine-/human-readable distinction.

2.4 Privacy policies under the EU and the U.S. law

Before discussing the legal requirements for privacy policies’ content and form, we want to highlight the different mindsets through which European and American academia approach the legal status of privacy policies. The main question is: should one treat a privacy policy as an information obligation (a mandated disclosure) or as a (boilerplate) contract? While the latter approach is common in the US, Europeans usually adopt a different perspective. In this subsection, we outline the reasons behind this difference and its consequences. In EU, consumer privacy is governed by the horizontal General Data Protection Regulation (GDPR) (Yeung and Bygrave 2022). It applies not only to online platforms but generally to all private and public

entities collecting and using personal data. Among many obligations, it requires data controllers and processors to fulfill transparency requirements, by publishing privacy policies (art. 12–14 GDPR). A failure to do so, or a divergence between what is stated in the privacy policy and what corporations actually do, can result in administrative fines issued by Data Protection Authorities in Member States (Gentile and Lynskey 2022). Within the GDPR logic, privacy policies are never sources of legal rights to process information, but solely information obligations (mandated disclosures). Indeed, legal rights need to be secured elsewhere, e.g., through consent or a necessity to perform a contract (art. 6 GDPR). Europeans—who otherwise regulate boilerplate contracts extensively (Reich and Micklitz 2014)—do not consider privacy policies as contracts and seldom, if ever, apply contract law rules to these documents.

In the United States, the situation is different alongside two dimensions: (1) the source of the obligation to publish privacy policies, and (2) the enforcement mechanism whenever corporations behave differently than stated in privacy policies. Regarding the former, in the United States, unlike in the EU, there is no federal statute governing online privacy, and thereby no federal statutory requirement to publish privacy policies (Yerby and Vaughn 2022). Instead, these matters are governed by several state-level statutes, with CalOPPA⁸ being the first one, and many others adopted in the recent years (Newell et al 2023). This said, it would be inaccurate to claim that there are no federal *rules* in the U.S. requiring to publish privacy policies. Such a requirement exists but it is not statutory in nature. It stems from decisions by the Federal Trade Commission, called by some scholars the “common law of privacy” (Solove and Hartzog 2014). Not publishing a privacy policy would be considered a deceptive omission by the FTC (Solove and Hartzog 2014).

Regarding enforcement, as the FTC has been initially slow to act, consumers started relying on contract law as a tool for policing corporations which behave differently from what is stated in privacy policies (Solove and Schwartz 2020). This led to a debate on whether these documents should be treated as contracts (Davis and Marotta-Wurgler 2019) or not (Solove and Schwartz 2020). If the answer is positive, federal and state court decisions concerning boilerplate enforceability—e.g., *Carnival Cruise Lines, Inc. v. Shute*⁹ (on choice of forum), *AT&T Mobility LLC v. Concepcion*¹⁰ and *American Express v. Italian Colors*¹¹ (on mandatory arbitration)—would apply to privacy policies. The problem of boilerplate’s justiciability gave rise to fascinating scholarly discussions (Sterkin 2004; Gibson 2017; Encarnacion 2019) which, however, we do not consider directly relevant to our argument.

In one sense, it is always possible (also within the EU law) to argue that a privacy policy gets incorporated, as a future-looking promise, into a contract. However, we are less interested in privacy policies as (potential) sources of corporate rights to process data and more in their function as transparency tools (mandated disclosures) meant to reduce information asymmetries. For this reason, in the remainder of this

⁸ California Online Privacy Protection Act, Cal. Bus. & Prof. Code §§ 22,575–22,579.

⁹ 499 U.S. 585 (1991).

¹⁰ 563 U.S. 333 (2011).

¹¹ 570 U.S. 228 (2013).

paper, we do not engage the contract law problems and treat privacy policies solely as information obligations.

3 Legal requirements for privacy policies: comprehensiveness vs. comprehensibility

The law, both in the European Union and in the United States, assigns the drafters of privacy policies a difficult task. On the one hand, privacy policies should be *comprehensive*, i.e., contain all the information relevant to a consumer pondering whether to use a service. On the other, they should be *comprehensible*, i.e., written in simple language and easy to understand for consumers.

In the United States, online consumer privacy is governed by the so-called “notice and choice” model (Richards and Hartzog 2018). As in the U.S. there exists no comprehensive federal statute concerning online data privacy, the model’s sources include the decisions by the Federal Trade Commission, called by some scholars the “common law of privacy” (Solove and Hartzog 2014) and a growing number of state privacy laws (Newell et al 2023). Within its logic, companies who wish to collect and use personal data should make privacy policies available to consumers (“notice”) who, in turn, can decide whether such data practices are acceptable to them (“choice”). Whether companies actually live up to their promises has been policed by both contract law (Bar-Gill et al 2017) and public enforcement by the F.T.C. (Solove and Hartzog 2014) and state Attorneys General (Citron 2016).

The latter can also issue guidelines. For example, in 2014, then-AG of California, Kamala Harris, stated that privacy policies “should provide a comprehensive overview of [...] practices regarding the collection, use, sharing and protection of personally identifiable information” and recommended that drafters of privacy policies “use plain, straightforward language. Avoid technical or legal jargon. Use short sentences” (Harris 2014).

In the European Union, the GDPR is the main instrument governing online consumer privacy (Yeung and Bygrave 2022). It requires that data controllers provide consumers with detailed information about their data practices,¹² yet do so in “in a concise, transparent, intelligible and easily accessible form, using clear and plain language.”¹³ The two requirements are in obvious tension.

Regarding comprehensiveness, under the GDPR’s transparency requirements, the *unit of information* that must be disclosed is a single *act of processing* of personal data. An act of processing means any action that a data controller takes regarding personal data.¹⁴ Thus, transparency at the stage in which personal data is collected requires the consumers to be informed, among others, about the following:

1. Categories of personal data concerned (e.g., an email address);

¹² The GDPR, arts. 13–14.

¹³ The GDPR, art. 12.1.

¹⁴ GDPR art. 4.1.

2. Purpose of processing for which each category of data is intended (e.g., issuing receipts)
3. Legal basis for each data processing¹⁵ (e.g., contractual necessity)
4. Storage period or, when it is not possible, the criteria to determine such a period (e.g., till account closure), and with regard to each data and purpose
5. Recipients or at least the categories of recipients (e.g., a cloud provider), as well as their
 - (a) Role (controller or processor)
 - (b) Purpose of sharing (e.g., data backup)
 - (c) Legal basis of sharing (e.g., legitimate interest)

This is quite a lot of information. Simultaneously, the GDPR requires this information to be presented in a concise and simple manner. The guidelines endorsed by the European Data Protection Board¹⁶—a body comprised of the Supervisory Authorities from all the EU Member States—provide some examples of how to square the two requirements:

Poor practice examples

“We may use your personal data to develop new services” (as it is unclear what the “services” are or how the data will help develop them);

“We may use your personal data for research purposes” (as it is unclear what kind of “research” this refers to) and;

“We may use your personal data to offer personalised services” (as it is unclear what the “personalisation” entails).

Good practice examples

“We will retain your shopping history and use details of the products you have previously purchased to make suggestions to you for other products which we believe you will also be interested in” (it is clear what types of data will be processed, that the data subject will be subject to targeted advertisements for products, that their data will be used to enable this).¹⁷

This looks clear and simple when presented as an isolated example. However, when one considers the sheer volume of data processed by online companies nowadays, the multitude of purposes for which different categories of data are used, or the range of entities with whom data is shared, a *comprehensive* privacy policy quickly becomes inflated. Quite possibly, the data controller from the EDPB’s example uses consumers’ shopping history for several other purposes and provides suggestions

¹⁵ Art. 6.1. GDPR lists six possible bases for processing: (a) consent, (b) contractual necessity, (c) legal obligation, (d) subject’s vital interest, (e) public task, or (f) controller’s legitimate interest. In the analyzed context, (a), (b), (c) and (f) are relevant. Whenever a data controller relies on (f), they should specify what legitimate interest they have in mind, to enable the proportionality assessment (Article 6(1)(f)).

¹⁶ Art. 29 WP, Guidelines on transparency under Regulation 2016/679, last Revised and Adopted on 11 April 2018, available at: <https://edpb.europa.eu/our-work-tools/our-documents/guidelines/transparency>. Hereinafter “The Transparency Guidelines.”

¹⁷ All the examples are taken from the Article 29 WP, Transparency Guidelines, p. 9.

based on other categories of data as well. Moreover, even this example does not meet all of the art. 13 GDPR requirements, as it is silent about the legal basis of this processing or the time period of data storage. Against this backdrop, the drive towards more comprehensibility (i.e., shorter and simpler policies) advocated for in scholarship (Calo 2011; Emami-Naeini et al 2021; Kelley et al 2009; Waldman 2018), needs to be re-assessed. At what point does the short and easy-to-read privacy policy become devoid of meaning?

We argue that a privacy policy should include enough information to comprehensively answer all reasonable questions consumers might have about data practices by service providers. This is in line with the overall philosophy of the informational self-determination paradigm pursued by the GDPR (Vivarelli 2020) and American privacy law (Reidenberg et al 2015). The idea, basically, is that a consumer will be well-positioned to decide whether to use a service and/or what data to reveal while doing so, upon knowing what data the company plans to collect, how it plans to use it, and with whom it plans to share it, as well as for what purposes. Hence, in the next section, we propose nine examples of questions that a consumer might reasonably be interested in. We will further use these questions to assess the comprehensiveness of real privacy policies and test our proposed format of a fully comprehensive privacy policy.

4 Assessing privacy policies: nine questions on data processing

With the rise of the gig-economy and in the aftermath of the COVID-19 pandemic, individuals' reliance on food delivery apps has grown (Raza et al 2023).

We consider them a gentle case study for the present research for several reasons. First, the type of data food delivery apps collect might, on the surface, seem banal, yet researchers have shown how dietary choices correlate with sensitive information like one's race (Marshall and Naumann 2018) or political convictions (DellaPosta et al 2015). Such data can also be beneficial for health insurers, interested in adjusting premiums of consumers developing (un)healthy dietary habits. Second, this is a competitive market, where one could imagine, at least theoretically, service providers competing on the privacy terms they offer (Marotta-Wurgler 2016). Third, for the purposes of our research, the catalogs of data categories these companies collect is wide, but more limited than those of data behemoths like social media, rendering the experiments interesting but feasible.

Hence, we imagine nine questions that a consumer could be interested in when considering whether to use a food delivery app, or choosing among different apps. This is not to say that these are all possible questions; rather, in our view, they well exemplify the kinds of information information consumers could be interested in having access to.

Q1: What data does Orderoo process about me?

Q2: For what purposes does Orderoo use my email address?

Q3: Who does Orderoo share my geolocation with?

Q4: What types of data are processed on the basis of consent, and for what purposes?

Q5: What data does Orderoo share with Facebook?

Q6: Does Orderoo share my data with insurers?

Q7: What categories of data does Orderoo collect about me automatically?

Q8: How can I contact Orderoo if I want to exercise my rights?

Q9: How long does Orderoo keep my delivery address?

In the following sections, we'll use these questions to assess: (i) the comprehensiveness of five actual privacy policies of food delivery companies, i.e., Deliveroo, DoorDash, Glovo, Just Eat, and Wolt; (ii) the ability of two LLMs to answer such questions based on the mock, fully-comprehensive, privacy policy we created; and (iii) test what would happen if the same LLMs were asked to answer these questions based on the actual, non-comprehensive, privacy policies.

5 Human evaluation of existing privacy policies (scenario 1)

In the first scenario, we tested whether legal experts can answer the selected questions, based on privacy policies currently in use. Our assumptions, supported by legal research in the area (Pollach 2011; Reidenberg et al 2015), were that privacy policies should be accessible and understandable by all users with or without a deeper understanding of the subject matter. Second, legal experts with knowledge of data protection laws and having read a substantial number of policies should ideally be able to answer all questions accurately. Third, we assumed that involving legal experts and not setting any time limits would allow us to overcome the usual limitations of analysing policies, i.e., their length, poor readability scores, and time limitations.

Our case study focused on the food delivery sector, therefore we chose five privacy policies from well-known service providers in this area: Deliveroo, DoorDash, Glovo, Just Eat, and Wolt. The latest versions of the policies were downloaded in English on 10 January 2024. The length of policies varied between 4379 words (Deliveroo) and 13,765 words (Glovo). Each policy was independently evaluated by two legal experts. The experts were required to answer the 9 questions, specifying for each of them whether the policy provided (a) unambiguous (clear) information, so that a definite answer could be given, or (b) ambiguous (impossible to know) information, so that no definite answer could be provided. The experts were also requested to add justifications for their evaluations, based on the analysis of the text.¹⁸ The evaluator agreement was equal to 100%.

Table 1 illustrates the findings—only two out of nine questions could be answered unambiguously. These two questions relate to whether the company shares data with insurers (Q6, clear in 4 out of 5 policies) and contact information for users to exercise their rights (Q8, clear in 5 out of 5 policies). Comparatively, these were the easiest questions to answer, since the relevant information was available in a single paragraph (as is usual).

¹⁸ Full details of both evaluations and can be found on GitHub <https://github.com/marcolippi83/comprehensive-privacy-policies>.

Table 1 Legal expert evaluation of privacy policies in the food delivery sector

Question	Just Eat	DoorDash	Wolt	Glovo	Deliveroo
Q1	X	X	X	X	X
Q2	X	X	X	X	X
Q3	X	X	X	X	X
Q4	X	X	X	X	X
Q5	X	X	X	X	X
Q6	✓	X	✓	✓	✓
Q7	X	X	X	X	X
Q8	✓	✓	✓	✓	✓
Q9	X	X	X	X	X

The majority of questions (7 out of 9) could not be answered with certainty, in other words, the information was ambiguous. The challenge of answering these questions stems from several problematic practices. We list the most common practices with examples from the policies below. Note that these issues often co-occur resulting in further reduction of clarity and comprehensiveness about the data handling practices.

The most common problem in the policies was the vagueness of the language used. Such language introduces uncertainty about what data is collected and how this data is used. Vague language introduced ambiguity in different ways—e.g., data categories were defined only at a high level ('personal data,' 'your data,' 'your information'), open-ended qualifiers were used such that no clear conclusions could be drawn from these explanations ('may,' 'could,' 'would,' 'as necessary'), examples ('such as,' 'for example') were used instead of providing clear and exhaustive lists for data types and processing practices. Consider the examples of such language practices taken from the real privacy policies below.

A4: "We process your personal data to the extent necessary and based on legitimate interest, for example, to: [...] market the Wolt Services to you or show you targeted or personalized advertisements through Wolt Services [...]" [Wolt].

A9: "We will only retain your information for as long as reasonably necessary to fulfil the purposes we collected it for, including for the purposes of satisfying any legal, regulatory, tax, accounting or reporting requirements." [Deliveroo].

A7: "We also may collect and store certain information about you and your device(s) automatically when you access or use our platforms and Services, including, without limitation, on our websites and engaging with our online chat functionality" [DoorDash].

Some of our questions asked for information about specific data types. For example, Q2 asked about the purposes for which the user's email is used, and Q9 required information on the retention period of the delivery address. To enable definite answers to these questions, the policies should specify purposes and retention periods for every data type collected. None of the policies included such specifications. See the examples below for typical phrasing used in the policies.

Table 2 Legal expert evaluation of privacy policies in the e-commerce sector

Question	Amazon	AliExpress	eBay	Shein	Zalando
Q1	×	×	×	×	×
Q2	×	×	×	×	×
Q3	×	×	×	×	×
Q4	×	×	×	×	×
Q5	×	×	×	×	×
Q6	×	✓	×	✓	✓
Q7	×	×	×	×	×
Q8	✓	✓	✓	✓	✓
Q9	×	×	×	×	×

A9: “Wolt does not store your personal data longer than is legally permitted and necessary for the purposes of providing the Wolt Services or the relevant parts thereof.” [Wolt].

A9: “Our retention of your personal information may vary based upon the nature of our relationship with you or, as applicable, the company or business with whom you are employed, associated, or connected (e.g., in the case of Merchant Personnel and Service Provider Personnel, etc.) and based upon several factors, including without limitation, the following: [...]” [DoorDash].

A2: “Glovo processes your personal data for: i) detect and investigate fraud and possible crimes committed against our Platform and all the users, [...]” [Glovo].

In our analysis of the privacy policies, we have not encountered a single policy that is compliant with all the GDPR requirements. In many instances, even legal experts are unable to answer the proposed questions with absolute certainty. This is due to the lack of both comprehensiveness and transparency. Moreover, our analysis showed that the quality and completeness of the information provided are not linked to the policy length.

To verify and argue for the general significance of our findings within the current privacy policy landscape, we selected and analyzed five additional privacy policies from the e-commerce sector, i.e., Amazon, AliExpress, eBay, Shein, and Zalando. We decided to focus on this domain since the 9 selected questions can be easily applied, allowing for a direct comparison of comprehensiveness and transparency with the food delivery sector. The selected privacy policies were downloaded on 7 October 2024. The length varied between 4533 words (Shein) and 20,512 words (eBay).

As in the previous case, each policy was independently evaluated by two legal experts, based on the same criteria.¹⁹ The evaluators’ agreement amounted to 100%. As shown in Table 2, the results are consistent with those of the first privacy policy

¹⁹ Full details of both evaluators and can be found on GitHub <https://github.com/marcolippi83/comprehensive-privacy-policies>.

set. Only 2 out of 9 questions, i.e., Q6 and Q8, could be answered. The highlighted evidence of vagueness, ambiguity and lack of required information is further confirming the outcomes of previous studies (Contissa et al 2018).

6 The legal proposal: fully comprehensive privacy policies

Our policy suggestion aims to create an informational environment where AI's full potential in assisting consumers can be realized. We propose that the law should require corporations to disclose *fully comprehensive* privacy policies, ideally in a standardized form, and relax the *comprehensibility* requirement. Such notices can be later processed by other parties to suit particular consumers' informational needs, particularly via automated means.

In this paper, we propose a format for *fully comprehensive* privacy policies. We suggest organizing the text by *data category*, i.e., devoting one separate paragraph for each type of data processed by a controller. These data categories should themselves be specific and unambiguous, i.e., for example, read “the amount of time you spend in the app from opening to order,” or “the type of food you search for,” etc., rather than “usage information.” The test for non-ambiguity is simple: if it is reasonably possible for a reader to ask “but what do you mean by category X?” that category should either be split into smaller categories or exhaustively defined. As an example of the latter, consider the excerpt of our mock privacy policy: “By payment information, we mean: bank account number (if you pay by bank transfer) or credit card number, expiration date, security code and billing postcode (if you pay by credit card), date and time, payment amount, PayPal email address or IBAN information.” To satisfy the exhaustiveness requirement, enumerations should *not* contain phrases like “such as,” “for example,” as these phrases can render the text ambiguous again.

Each paragraph should contain the following information: (A) data category, (B) source, (C) purpose of processing, (D) purpose explanation, (E) legal basis, (F) legal basis explanation, and (G) storage period. Further, in case the data category will be shared with third parties, the second part of the same paragraph should contain information on (I) the recipient's identity, (II) their role (processor and controller), (III) the purpose of sharing, (IV) legal basis for sharing for each specific purpose. We illustrate how this could look like in Appendix A.

Note that this policy has been designed, first and foremost, with *human expert readers* in mind. The goal was to create a document that would enable a person with technical knowledge of law and resources like time and cognitive capacity, to unambiguously answer all questions. This is not to say that we expect consumers to spend time reading these long documents. Rather, our hypothesis was that, based on a such a document, LLMs should also be able to provide answers while retaining the ability of humans to always verify them.

7 LLMs and the mock privacy policy (scenario 2)

In the second scenario, our goal was to assess to what extent LLMs were able to correctly answer user questions regarding the contents of fully comprehensive privacy policies. We asked the nine questions introduced in Sect. 4 with respect to the Orderoo mock privacy policy described in Sect. 6.

The three initial questions, as well as Q8, are rather straightforward and aimed to test the model's ability to retrieve the relevant information. The answer to Q1 can be found at the beginning of each paragraph of the mock policy, while the answers to Q2, Q3 and Q8 are contained in a single paragraph. Q4, Q5, Q6, Q7 and Q9 are more complicated. The answers to Q4, Q5, and Q7 are spread throughout the entire document. Q6 requires a negative answer, i.e., the policy does not mention any data sharing with insurers. It is mostly aimed at testing potential hallucinations. Q8 is a question where a user would ask about a data category that could correspond to two different data categories in the document (i.e., contact email and business address); we wanted to see if this would confuse the LLMs.

7.1 Experimental setup

We provide the tested LLMs with the full text of the policy, and the list of questions, numbered from 1 to 9. Since the performance of all LLMs is strongly dependent on the way in which prompts are engineered (White et al 2023), we test the following two prompt variations:

Prompt 1 (short): In answering the questions please rely solely on the information included in the text and not your knowledge from other sources; please read the document carefully and mention everything, do not omit any information included in the text; please do not shorten or simplify the answers.²⁰

Prompt 2 (Chain-of-Thought): In answering the questions please rely solely on the information included in the text and not your knowledge from other sources. Do not omit any information included in the text, and do not shorten or simplify the answers. Please reason step by step.²¹

The prompts aim to give precise instructions, to mitigate hallucinations (i.e., making answers up) (Azamfirei et al 2023; Bang et al 2023), unwarranted simplification (i.e., listing only some answers, preceded by phrases like “for example” or “such as”), and unwarranted text generation (i.e., adding information that the user did not ask for, e.g., advice that one should carefully read the laws, etc.).

As for the candidate LLMs to be evaluated, we performed many preliminary experiments. We tested the following models: GPT-4, Llama-7B, Llama-13B, Llama-70B, Mistral-7B, and Momo-70B. We observed a general behaviour with most of the considered models: except for GPT-4, all the other LLMs struggled with

²⁰ Experiments with Prompt 1 were carried out in January 2024.

²¹ Experiments with Prompt 2 were carried out in October 2024.

dealing with a large context, as the mock policy is 6880 words long. GPT-4 is, in fact, the only model, among the ones we tested, having a default context window up to 128k tokens.

We are aware that LLMs' performance may also vary based on hyper-parameters adjustments. However, we did not engage in LLM fine-tuning, which would have required a completely different methodology, with the need for a training data set. We believe this is beyond the scope of this paper, although certainly worth studying in the future.

Our preliminary analysis showed that the results of GPT-4 stand out significantly for their quality. The results obtained with the three Llama versions are similar (though with a much lower quality than GPT-4 results), but slightly better for the smallest model (7B), the larger ones being more prone to summarization and hallucination. As for Mistral-7B, we were not able to get valuable answers, again probably due to the very large input context. The answers obtained with Momo-70B also showed significant mistakes. While we provide the full transcripts of all these preliminary experiments in our Github repository, in the following we will analyze in details only those obtained with GPT-4 and Llama-7B.²² Nevertheless, we remark that obtaining better results with all the LLMs could indeed be possible, with a combination of model calibration, prompt engineering, and fine-tuning. Our goal is not to exhibit the *winning* model, but rather to show that adequately instructed LLMs can successfully accomplish this task.

For both GPT-4 and Llama2-7B, we repeated the test five times to assess variability across different runs, as LLMs are stochastic generative models. All the answers were manually checked by legal experts.

7.2 GPT-4 results on mock policy

For GPT-4, the very high quality of the results allowed us to employ standard classification metrics that are typically used for question-answering, information retrieval, and machine learning in general. In fact, we could easily identify two main categories of mistakes made by the LLM: (i) false positives (FP), where the reply contains incorrect information, though based on the document and (ii) false negatives (FN) where the reply misses information contained in the document. As we shall see later, incorrect information not based on the document is classified as hallucination. Given also the number of correct elements present in each reply (true positives, TP), one can introduce precision (P) and recall (R), as follows:

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad (1)$$

²² We remark that GPT-4 is a closed model, whereas Llama-7B is open. As additional information, we report that the overall cost of our experiments, including preliminary ones, with GPT-4 was below 3 dollars.

Just to give an example, Q4 asks for the types of data that are processed on the basis of consent, also asking for purposes. The correct answer should contain 5 categories: default delivery address, payment information, geolocation, contacts, and dietary restrictions. Therefore, in case the answer provided by GPT-4 for a single run contains only 3 out of 5 categories, we count 2 FNs, and we report a recall equal to 60.0. If, instead, the answer contains an additional category which was not among the 5 correct ones, we report a FP, thus leading to a precision equal to 80.0.

Table 3, reports the results obtained with GPT-4, relatively to Prompt 1 (short) and Prompt 2 (Chain-of-Thought).

Prompt 1 (short). As regard Prompt 1, for six questions (Q1-Q2-Q3-Q6-Q8-Q9) the answers are perfect on all five runs, with neither a FP, nor a FN. The three remaining questions produced a few mistakes, though both precision and recall are still very high. As for Q4, two runs out of five produced one FP each, and a single run produced one FN. Regarding Q5, all runs except one contained at least one FN, whereas two runs also had a FP, with one run reporting the correct reply. Q7 was the most problematic question in terms of recall, since every run consistently showed at least two FNs: more precisely, GPT-4 consistently missed “contact list details” and “call recordings” as data categories that are automatically collected by Orderoo. We decided to further investigate this point, and we asked GPT-4 explicitly (in two subsequent questions) whether such data categories are in fact automatically collected: the answer of GPT-4 this time was correct. We argue that this second way of phrasing the question (i.e., in terms of a yes/no question rather than asking for a list of possibilities) is simpler to answer. This is not surprising, since the performance of LLMs is known to be strongly dependent on the prompt.

To illustrate the performance of GPT-4 Prompt 1 on the mock policy, consider the following examples. As regard the question on who does Orderoo share the user’s data with (Q3), the correct answer should include four entities—delivery persons, vendors close to you, Facebook, and Google. GPT-4 correctly answered the question in all 5 runs with very minor variations between the runs.

Table 3 Results obtained by GPT-4 on the nine questions asked on the Orderoo mock privacy policy

	Prompt 1 (Short)			Prompt 2 (CoT)		
	Precision	Recall	F1	Precision	Recall	F1
Q1	100.0	100.0	100.0	100.0	100.0	100.0
Q2	100.0	100.0	100.0	94.0	100.0	96.9
Q3	100.0	100.0	100.0	100.0	100.0	100.0
Q4	96.7	92.0	94.3	96.7	72.0	82.5
Q5	94.4	90.0	92.1	100.0	92.0	95.8
Q6	100.0	100.0	100.0	100.0	100.0	100.0
Q7	100.0	78.2	87.8	100.0	81.8	90.0
Q8	100.0	100.0	100.0	100.0	100.0	100.0
Q9	100.0	100.0	100.0	100.0	100.0	100.0

Precision accounts for the false positives and recall for the false negatives. Results are averaged over 5 runs

Prompt 1, Run 1: “Orderoo shares your geolocation with delivery persons (for order delivery), with vendors close to you (to facilitate their own marketing), with Facebook, and with Google (both for ad delivery and third-party marketing).”

Prompt 1, Run 3: “Orderoo shares your geolocation with delivery persons for order delivery, vendors close to you for their own marketing, Facebook and Google for ad delivery and third-party marketing.”

As for the errors, GPT-4 Prompt 1 did not produce any hallucinations. The few false negatives were related to missing one or two items from the list. Similarly, the false positives were related to the addition of one or two items to the list. Consider the following answers to the question about what data does Orderoo share with Facebook (Q5). The correct answer should include 10 data types, see the comparison between the synthesised list from the mock policy and the GPT answers.

Orderoo: “Data shared with Facebook: default delivery address, each delivery address, details of your order (what you buy), the price you pay, ratings of orders, geolocation, food interests, content preferences, system language, and country of residence.”

Prompt 1, Run 2: “Orderoo shares the following data with Facebook: your default delivery address, each delivery address, for the purpose of ad delivery and third-party marketing, your geolocation, the details of your food interests, your ratings of the orders, your system language, your country of residence, and your content preferences for the purpose of personalizing advertising. [FN: GPT missed ‘details of your order’ and ‘the price you pay’.]

Prompt 1, Run 4: “Orderoo shares the following data with Facebook: email address, default delivery address, each delivery address, order details, ratings of your orders, your geolocation, content preferences, food interests, system language, country of residence.” [FN: GPT missed ‘the price you pay.’ [FP: GPT added. ‘email address.’]

Overall, GPT answered the majority of questions correctly (33 out of 45 questions) and followed the prompt instructions carefully, i.e., it used only the information available in the text and did not shorten or simplify the answers.

Prompt 2 (Chain-of-Thought). As regard Prompt 2, for five questions (Q1-Q3-Q6-Q8-Q9) the answers are perfect on all five runs (29 out of 45 questions). In comparison with Prompt 1, the results obtained with Prompt 2 are slightly worse for Q2 and Q4. Concerning Q2, GPT-4 produced 3 FPs in one run. Differently from what was observed with Prompt 1, Q4 was the most problematic question, since all runs produced one or more FN, and a single run produced one FP. In contrast, for Q5 and Q7, Prompt 2 slightly outperforms Prompt 1. Note that, for Q7 the issue related to FN persists, since the model produced the same 2 FNs in each run. The most relevant difference compared to Prompt 1 is that, Prompt 2 produced an inference for Q2 in a single run. As further detailed in Sect. 7.3, we treat as inferences answers indicating that LLMs have drawn conclusions based on the policy, which may or may not be true. In particular, in addition to the

correct purposes for using the user email address, GPT-4 added the need “To determine trends in purchase behavior” which seems to follow from the profiling purpose. Finally, Prompt 2 also displayed several repetitions (REP), i.e., the model reported the same items multiple times for Q2 and Q5 in a single run. For example, when answering Q5 on what data are shared with Facebook, the output included “food interests” twice.

Orderoo: “Data shared with Facebook: default delivery address, each delivery address, details of your order (what you buy), the price you pay, ratings of orders, geolocation, food interests, content preferences, system language, and country of residence.”

Prompt 2, Run 2: “[...] Details of your food interests for ad delivery and third-party marketing, [...] Food interests for ad delivery [...]

To assess these results against another LLM, we report the same prompts and questions used with Llama2-7B.

7.3 Llama2-7B on mock policy

The results obtained with Llama2-7B are significantly worse than those achieved with GPT-4, relatively to both Prompt 1 (short) and Prompt 2 (Chain-of-Thought). As shown in the following, Llama2-7B replies were much more complex to analyze even for humans, since very often it tries to summarize the content of the policy, and sometimes it also hallucinates, i.e., it provides (incorrect) information not based on the document. It is not easy to frame such mistakes into the two plain categories of FPs and FNs, necessary to compute precision and recall. Thus, we decided not to report for Llama2-7B the same quantitative evaluation produced for GPT-4 and measured the correct and wrong answers in a different way (the same assessment criteria were used in Sect. 8).

In particular, correct answers included all the true positives and true negatives. We distinguish five types of errors: (1) false negative (FN) when the answer is missing information present in the policy, (2) false positive (FP) when the answer includes information that is not relevant for answering the question, but can be found in the policy, (3) summarization (SUM) when the answer indicates that LLM has summarised information in the policy, (4) inference (INF) when the answer indicates that LLMs made inferences based on the policy, which may or may not be correct, and (5) hallucinations (HAL) when the answer includes information that is not in the text. Each answer was treated as one unit for counting the errors and one answer could include multiple errors.

Prompt 1 (short). With regard to Prompt 1, Llama2-7B provided only one correct answer for the 45 questions (9 questions for 5 runs). Evaluation of Llama2-7B under the above-mentioned criteria shows a very high count of missing information (FN=91%) and a staggering frequency of hallucinations (HAL=62%), i.e., answers not based on the analyzed policy. Prompt 1 answers also include moderately frequent instances of summarization (SUM=27%), false positives based on the

information present in the policy but irrelevant for the question at hand (FP = 38%), and inferences (INF = 13%).²³

While the missing information can be identified fairly easily, other types of errors are more complex and introduce potentially more problematic issues. Consider, for example, answers to the question about how long Orderoo keeps the delivery address (Q9) with the correct answer of “until you delete your account, plus three months”:

Prompt 1, Run 3: “Orderoo keeps your delivery address for 18 months after the last order.”

Prompt 1, Run 4: “According to the document, Orderoo keeps your delivery address for a period of 18 months after the last order placed. After this time, your data will be deleted or anonymized, unless you have given consent for longer retention.”

Both answers are not based on the information found in the text, which we considered as hallucinations. Moreover, Llama2-7B’s answers do not indicate any uncertainty about the information given, exemplified by the following phrase “according to the document.”

Consider another example based on the question about the contact details to exercise the user’s rights (Q8) with the correct answer of “By email: dpo@orderoo.com or privacy@orderoo.com. By mail: 1 Name Street, 40,121, Bologna, Italy”:

Prompt 1, Run 1: “[...] you can contact their customer service team at [insert contact information]. Please note that you may need to provide proof of identity to verify your request.” Prompt 1, Run 2: “[...] * Email: [info@orderoo.com] (mailto: info@orderoo.com) * Postal address: Orderoo, 123 Main St, Anytown, AB12 3CD.”

Llama2-7B Prompt 1 failed to answer this question correctly in all five runs. As demonstrated in the examples above, the answers included missing information and hallucinations. In addition to the errors, we also observed interesting patterns in the different runs. For example, in run 3, answers included emojis, and in run 5, Llama used quotation marks in the answers, misleading the reader that it was quoting the policy’s text directly when it was not.

Prompt 2 (Chain-of-Thought). In comparison with Prompt 1, Prompt 2 performed worse. First of all, the model is more prone to include unrequested summarisation of the entire privacy policy, before answering the queries. Summarisation occurred three times in two out of five runs. Consider the following example:

Prompt 2, Run 5: “Orderoo processes data about you for the following purposes: * To provide the Orderoo service to you: This includes processing your orders, payment details, and communication with you regarding your order.
* To improve our services and products: Orderoo may use your data to ana-

²³ Full documentation of the error analysis can be accessed on GitHub <https://github.com/marcopolippi83/comprehensive-privacy-policies>.

lyze customer behavior and preferences, which helps us to improve our menu, services, and marketing efforts. * To send you marketing communications: Orderoo may use your email address to send you newsletters or other marketing messages. These messages may include special offers, promotions, or updates on new products or features. You can opt-out of receiving these messages at any time by following the unsubscribe link included in each message. * To personalize your experience: Orderoo may use your data to tailor our services and marketing efforts to better suit your needs and preferences. For example, we might show you personalized recommendations based on your order history or location. * Compliance with laws and regulations: Orderoo must comply with applicable laws and regulations, including those related to data protection. This means that we may need to disclose your data in response to a legal request from government authorities or law enforcement agencies.”

Secondly, the model skips answering questions completely 9 out of 45 questions in four runs, while producing a disclaimer to justify its behavior in two runs. Consider the following example, where the model produced a disclaimer after answering Q6 and then failed to answer from Q7 to Q9:

Prompt 2, Run 1: “I apologize for any confusion, but I cannot provide an answer to your questions as they are based on a fictional document that does not exist. The questions you provided are not real and therefore, I cannot rely solely on the information included in the text to answer them. I’m just an AI assistant trained to provide accurate and helpful responses, but I cannot create fake documents or answers to questions that do not exist. If you have any other questions or requests, please feel free to ask, and I will do my best to assist you.”

Overall, Llama2-7B Prompt 2, answered 36 out of 45 questions, where only two answers were correct. Thus, we can analyze errors only for these answers, making it more difficult to compare the results with Prompt 1. In particular, Prompt 2 produced 92% FNs, 42% FPs, summarization (SUM 28%), and hallucinations (HAL 64%). Notably, the percentage of inferences (INF 53%) is significantly higher compared to Prompt 1. Consider the following examples of errors produced with Prompt 2. As regard Q5, it seeks to extract the data shared by Orderoo with Facebook.

Orderoo: “Data shared with Facebook: default delivery address, each delivery address, details of your order (what you buy), the price you pay, ratings of orders, geolocation, food interests, content preferences, system language, and country of residence.”

Prompt 2, Run 3: “[...] Location data (e.g., latitude, longitude) for the purpose of providing location-based services and personalizing the user experience. [...]”

Note that, the output includes not only the categories of data, among other “location data,” but also a specification of what could be considered as geolocation information: “e.g., latitude, longitude”. While this inference (INF) might be correct, it violates the prompt’s restriction of reporting only what is in the

privacy policy. Moreover, as noted above, even if an inference is valid from a logical perspective, the conclusion may be false.

Llama2-7B also produced text summarization in answering the questions, even if prompts explicitly forbid it. Consider the following example relative to Q6, on what data are processed on the basis of consent, and for what purposes.

Orderoo: “Default delivery address, for delivery address suggestion; Payment information, storing your payment information for future payments; Geolocation, delivery address suggestion; Contacts, to synchro- nize the contact list with the Orderoo app; Dietary restrictions and food allergies, highlighting the potential allergens among the food ingredient descriptions, and sharing with vendors.”

Prompt 2, Run 5: “Types of data processed on the basis of consent: personal data. Purposes: providing services and improving services.”

From the example above, it is clear that, rather than reporting all the data categories and the related purposes of processing, Llama2-7B groups all the information into “personal data”, also summarizing the mentioned purposes under the broader umbrella of “providing and improving services.”

7.4 Discussion of scenario 2 results

The performance of GPT-4 is really good under at least two metrics. First, on the meta-level, the fact that the answers given to the 9 questions so closely corresponded to the content and form of the correct answers—making it possible to compute precision and recall—is undoubtedly a success of the model. Note how the use of these standard metrics, for reasons presented above, was impossible in a straightforward manner for Llama2-7B (as it would have produced low-single-digit results, additionally missing the problem). Second, GPT-4 Prompt 1 scored really high: it answered 6 out of nine questions flawlessly, with F1 of 94.3 in Q4, 92.1 in Q5, and 87.8 in Q7. This is a very promising result from the law and consumer policy perspectives. Third, the comparison between short and chain-of-thought prompts for GPT-4 and Llama2-7B suggests that, for question-answering and information extraction tasks, step-by-step reasoning offers no advantages and can even reduce the performance. Even though the majority of the (pre-)tested LLMs, i.e., Llama-7B, Llama- 13B, Llama-70B, Mistral-7B, Momo-70B, did not manage to produce results satisfactory from the consumer policy point of view, we want to highlight that (i) they are in the early stages of their development; (ii) the prompts and settings we used were relatively straightforward. The fact that GPT-4 is capable of performing these tasks well indicates that work on other LLMs might render them so as well.

The really good performance of GPT-4 in the experimental scenario, i.e., answering questions about the fully comprehensive privacy policy, suggests that our policy proposal should be seriously considered by lawmakers or even privacy-sensitive corporations acting on their own motion. This proposal challenges the conventional wisdom that privacy policies should be made shorter rather than longer and simpler rather than more complex (Calo 2011; Emami-Naeini et al. 2021; Kelley et al 2009; Waldman 2018). Even though these proposals responded to the real trade-off

existing in the world of human-only readers, we argue that with the changing socio-technological reality, the law should reconsider the assumptions made about the audiences of privacy policies, and thereby start requiring well-structured, fully comprehensive texts that LLMs like GPT-4 can work with.

8 LLMs and real privacy policies (scenario 3)

Scenario 2 demonstrated that GPT-4 can accurately answer questions based on a fully comprehensive privacy policy, especially under Prompt 1 (short), while Llama2-7B produced only partially correct answers under both prompts. In scenario 3, we test the capabilities of these two models in answering questions, based on real privacy policies. The goal is to assess what results the two LLMs would produce when working with suboptimal source material. Given the better performance observed with the short prompt for both GPT-4 and Llama2-7B, in this scenario we decided to focus on Prompt 1. We also choose to limit this experiment to the five food delivery policies analyzed by humans in Sect. 5 (i.e., Deliveroo, Door-Dash, Glovo, Just Eat, and Wolt), excluding the e-commerce sector. The restricted focus is justified in light of the highly similar results obtained for the two market areas, showing the same problems from both a quantitative and qualitative perspective. In both sectors, the experts assessment clearly underscored the impossibility to unambiguously answer the nine questions. Including the e-commerce documents would have likely yield redundant results, offering little new insight into the LLMs' performance. Conversely, the restraint focus allows for a more efficient and targeted analysis, without compromising the validity of our findings.

The answers by GPT-4 and Llama2-7B were evaluated by legal experts. They compared the answers with relevant excerpts from the policy, indicating whether it was correct or contained one of the errors explained below. Legal experts were also required to provide a rationale for their evaluations.²⁴

LLMs had to provide 45 answers in total, 9 answers for each of the 5 policies. Surprisingly, GPT-4 provided only 9 correct answers, while Llama2-7B succeeded only in one instance. As for the error rates, false negatives occurred in 47% of GPT-4's answers and 64% of Llama2-7B's answers, indicating that a lot of relevant information was missed. False positives were found in 20% of GPT-4's answers and 58% of Llama2-7B's answers, accounting for information presented in the policy but irrelevant for answering the question. Hallucinations were not found in any of the GPT-4 answers. In contrast, Llama2-7B was prone to hallucinations with 38% of answers, including information that is not present in the privacy. GPT-4 summarized information more often than Llama2-7B, i.e., 24% answers of GPT-4 vs only 11% of Llama2-7B. Lastly, both models inferred information when answering the questions—GPT-4 in 33% and Llama2-7B in 31% cases.

²⁴ Full documentation can be found on our GitHub <https://github.com/marcolippi83/comprehensive-privacy-policies>.

Generally, GPT-4's answers were longer and more detailed than Llama2-7B's. Recall from the human analysis of the policies that only very few questions could be answered unambiguously. We did not ask the LLMs to answer about the clarity of the information in the policies, but to only use the text in providing answers. However, Llama2-7B was more likely to omit the uncertainty which is present in the selected documents. It is worth considering some examples of the errors more closely.

Table 4 includes the three types of answers to question 9—the original policy text, GPT-4 answer and Llama2 answer. Glovo's policy states that the maximum period of keeping the user's data is 15 years. It did not explicitly state that the delivery address is kept for this period therefore GPT's answer was treated as an inference. In contrast, Llama's answer was not based on the policy's text at all and was marked as a hallucination.

For the analysis of Q3 about sharing geolocation (see Table 5), we assumed that the mention of sharing 'address' can be equated to 'geolocation.' GPT-4 only mentioned Dashers as data recipients and missed the other entities with whom data is shared, i.e., merchants or other users (FN). Llama2-7B mentioned some of the correct recipients (delivery partners (Dashers) and restaurants (Merchants)). It also added other recipients which are mentioned elsewhere in the text, even though they were not relevant for this question (i.e., marketing and advertising partners, service providers). Moreover, the model's disclaimer about not sharing the data "in the event of a merger, acquisition, or sale of assets" is in direct contradiction to what the policy states "We may disclose your personal information to third parties in order to participate in, or be the subject of, any sale, merger, acquisition, restructuring, joint venture, assignment, transfer or other disposition of all or any portion of our business, assets or stock (including in connection with any bankruptcy or similar proceedings) or other similar transaction(s), in which case we may disclose your personal information to prospective buyers, sellers, [...]." Such hallucinations are extremely dangerous, since the answer is not just inaccurate but completely false.

Table 4 Comparison on Glovo, GPT-4 and Llama2 answers

Question 9	How long does Glovo keep my delivery address?
Glovo's PP	Glovo shall retain your data for the duration of the contractual relationship and, after this has come to an end, for the period established by law for filing or defending the appropriate legal actions. This is established at a maximum of fifteen (15) years in order to comply with Glovo's legal obligations [...]
GPT-4 (short prompt)	Glovo keeps your delivery address for the duration of the contractual relationship with you. After the end of the relationship, they keep it for the period established by law to comply with legal obligations and defend against legal actions, which is a maximum of fifteen years. The actual retention period may be shorter depending on the legal provision applicable to each purpose of data processing, as detailed in Annex I of the privacy policy. [INF]
Llama2-7B (short prompt)	According to the document, Glovo keeps your delivery address for a maximum of 1 year after the completion of the delivery. After this period, your data will be deleted. [HAL]

Table 5 Comparison on Doordash, GPT-4 and Llama2 answers

Question 3	Who does DoorDash share my geolocation with?
DoorDash PP	iii. Disclosing Personal Information [...] b. Dashers, Merchants or Other Users [...] To help provide you with our delivery services we provide your information, including limited personal information (e.g., first and last names, address, phone number (however, your phone number is masked to protect your privacy)) and order details, to the Dashers who complete your requested deliveries
GPT-4 (short prompt)	DoorDash may share your geolocation data with Dashers to facilitate delivery of your orders. They may also use geolocation data to help you select your delivery address, to provide more accurate deliveries, to provide recommendations, and to enhance your user experience. [FN, INF]
Llama2-7B (short prompt)	DoorDash shares users' geolocation data with the following parties: * Third-party delivery partners (e.g., drivers and couriers) * Restaurant partners for delivery (to enable pickup and delivery orders) * Marketing and advertising partners (for targeted promotions and offers) * Service providers (e.g., data analytics and fraud prevention companies). It's worth noting that DoorDash may also share users' personal information with third parties in the event of a merger, acquisition, or sale of assets. [FP, FN, HAL]

Table 6 shows how GPT correctly identified “technical data about your equipment, browsing activities and patterns” and summarized the specifications of data provided by social media sites. The only category of data correctly identified by Llama is “device information.” All the other data were missing (FN). Moreover, the list provided by Llama included data categories from the policy which were not specified as collected automatically (FP).

Table 6 Comparison on Just Eat, GPT-4 and Llama2 answers

Question 7	What categories of data does Just Eat collect about me automatically?
Just Eat	Personal data that we collect automatically: [...], we may automatically collect technical data about your equipment, browsing activities and patterns. [...] We may allow you to use social media sites to create your account or link your account to the relevant social media site. These social media sites may provide us with automatic access to certain personal data they hold about you (e.g. content viewed by you, content liked by you, and information regarding the advertisements you have been shown or may have clicked on)
GPT-4 (short prompt)	Categories of data Just Eat collects about you automatically may include technical data about your equipment, browsing activities, and patterns. They collect this data through cookies, server logs, and other similar technologies. They may also collect data from your use of social media sites to create or link your account, which may give them access to personal data held by those sites. [SUM]
Llama2-7B (short prompt)	Just Eat collects the following categories of data about you automatically: * Device information (e.g., IP address, browser type, operating system) * Location data (e.g., geolocation, location history) * Order history and details (e.g., orders placed, itemized list of delivered food) * Payment information (e.g., payment method, billing address, payment history) * Personal information (e.g., name, email address, phone number) [FP, FN]

Table 7 Comparison on Glovo, GPT-4 and Llama2 answers

Question 8	How can I contact Glovo if I want to exercise my rights?
Glovo	You may exercise your rights free of charge at any time using the form available on our main website glovoapp.com »Contact us»Policies, or in the App, entering to: Help»Not related to an order»Policies. To exercise your rights click HERE [hyperlink]. If you have any questions, you can write to us at gdpr@glovoapp.com
GPT-4 (short prompt)	To exercise your rights, you can contact Glovo using the form available on their main website glovoapp.com »Contact us»Policies or in the App by going to Help»Not related to an order»Policies. You can also send an email to gdpr@glovoapp.com . [FN]
Llama2-7B (short prompt)	If you want to exercise your rights as a data subject, you can contact Glovo at the following address: Glovo Customer Support Av. Francisco de Goya, 21–25, 28,036 Madrid, Spain You can also reach out to Glovo's Data Protection Officer (DPO) at dpo@glovo.com . [HAL]

Human analysis of the selected policies showed that the question about the contact information could be answered unambiguously for all five policies. Table 7 shows that GPT-4 included correct details, with a minor omission of the possibility of contacting Glovo through an online form. In contrast, Llama2-7B provided false contact details. It is a pattern for Llama2-7B for this question: it failed to provide the correct contact details for all companies. An interesting pattern concerns the Llama2-7B reuse of the same data types for answering questions, regardless of the privacy policy and its text. For example, ‘payment information’ was included in 20 out of 45 answers, where only 8 such mentions were correct. It raised a suspicion that Llama2-7B might be using information learnt from typical privacy policies, rather than basing the answers on the submitted text.

To summarize, GPT-4 provided significantly more accurate and more relevant answers than Llama2-7B. However, when we compared the LLMs answers based on the mock policy and the actual once, it appeared that incomplete or incorrect answers were mostly provided where the policy did not include clear and comprehensive information. In such cases, false positives and hallucinations may mislead users over data practices and their rights.

We encountered some challenges when analysing the answers provided by LLMs. First, it was not possible to quantify the errors at the phrase level, since it was often unclear what had caused inaccuracies in the answers. Second, summarized text and inferences often could not be linked to the corresponding portions of the policy.

9 Conclusion and future work

In a world of human-only readers, a trade-off exists between comprehensiveness and comprehensibility: only privacy policies too long to be humanly readable could precisely describe the intended processing. This has led both scholars and policy-makers to put the emphasis on the *comprehensibility* of privacy policies. As a result, the existing policies, even when considerably long and complex, fail to provide

meaningful information. We argue that making such documents *fully comprehensive* is a pro-consumer move, provided that automated means for question-answering are available.

For this purpose, we designed an experimental setting including all of the following. First, to assess the information content of current privacy policies, two legal experts were independently required to answer nine questions based on five privacy policies in the food delivery sector. Results show that only 2 out of 9 questions could have a precise answer; the remaining could not be answered with certainty, due to vague or missing information. To verify the general significance of our finding within the current privacy policy landscape, we selected and analyzed five additional documents from the e-commerce market area. The results are consistent with those in food delivery, thus confirming the impossibility to unambiguously answer the nine questions.

Second, to remedy this informational gap, we designed a format for a fully comprehensive privacy policy and drafted a mock policy, from which all the required and necessary information (in particular to answer the 9 questions) could be easily elicited by human readers (with the time necessary to read them). The idea here is that, if the information necessary to answer consumers' question is present in the policy, not only human experts would be able to answer them but also LLMs could do so quickly and accurately.

Third, we submitted the mock policy to two LLMs, Llama2-7B and GPT4. To this end, we engineered appropriate prompts, (a short one and a Chain-of-Thought prompt), in which the mentioned questions were supplemented by a set of directives. We ran the test with both prompts five times for each LLM. GPT-4 Prompt 1 (short) stands out significantly, providing satisfactory answers to most questions: in 6 cases the answers are perfect, with neither false positives nor false negatives; the answers to the three remaining questions contain a few mistakes, though both precision and recall are still very high. GPT-4 followed the prompt instructions very precisely, without shortening or simplifying the answers. GPT-4 Prompt 2 performed slightly worse, answering perfectly to 5 out of 9 questions. New errors emerged in two runs, i.e., inferences and repetitions of the same text multiple times. Nevertheless, GPT-4 overall performance remains impressive. Conversely, the performance of Llama2-7B's is significantly worse, providing only one correct answer out of 45 questions (9 questions for the five runs), in the case of Prompt 1. In many instances, its answers do not accurately reflect the content of the policies: delivering partial and simplified accounts or even hallucinations (62% of cases). Prompt 2 (Chain-of-Thought) introduced further deviations from the given instructions. In particular, answers were prone to include unrequested summarization and skip answering questions. These results show that when given fully informative privacy policies, GPT-4 can (and other LLMs arguably could, in the future) provide consumers with meaningful and precise answers.

Fourth, we tested the LLMs' question-answering performance on the five food delivery privacy policies. Limiting the experiments to this sector, excluding the e-commerce, is justified in light of the strong similarity of the results in the two market areas, from both a quantitative and qualitative perspective. Indeed, as noted in Sect. 8, while including the e-commerce documents would not have offered new insights into the LLMs performance, the restraint focus allowed for a more efficient analysis, without compromising the general findings.

We submitted the selected documents to both GPT-4 and Llama2-7B. Given the higher performance obtained by the LLMs with Prompt 1 (short) in the mock policy analysis, we decided to run this experiment only with Prompt 1. Also in this scenario, GPT-4 provided more accurate and relevant answers compared to Llama2-7B. However, when compared the LLMs performance on the mock policy and the real policies, in the latter case the results were significantly worse: GPT-4 provided correct answers to 9 out of 45 questions, while Llama2-7B gave only one correct answer. When faced with vague and incomplete information, both GPT-4 and Llama2-7B produced a high number of FN and FP, summarising or omitting relevant content, or inferring new information. Consistently with what we observed in the mock policy scenario, hallucinations were not found in any of the GPT-4 answers, while Llama2-7B hallucinated in the 38% of cases. It clearly suggests that the low performance is due to the poor drafting of the analyzed policies.

The novelty of these experiments lies in showing that, in analyzing well-structured, fully informative privacy policies, LLMs can provide consumers with meaningful and precise answers. The good performance of GPT-4 suggests that this is a promising path. Hence, our results not only reveal the potential benefits of LLMs in legal applications, but also demonstrate that in a world where LLMs are available, the law can and should change.

In the future, we plan to expand our experiments by: (a) drafting more extensive and diverse fully-informative privacy policies, (b) testing additional LLMs; (c) linking the LLMs' answers to the relevant portions of the text, as well as (d) automating the quality control over such answers. Further scholarly research on LLMs' capacities will undoubtedly occur alongside further development of these models. Hopefully, the synergy between the two strands of research will lead to a world in which consumers can easily obtain answers about how corporations use their data.

Until today, the two requirements of comprehensiveness and comprehensibility in privacy policies, could not be jointly implemented, making it legally acceptable to settle for a trade-off. Our contribution shows this is no longer the case: it is possible to fully achieve both goals. However, for this to materialize, efforts on the computer science side and a rethinking on the lawmakers' side need to take place. The goal of this paper was to point these communities in the promising, albeit still heterodox, direction of developing *fully comprehensive* privacy policies to be analyzed by humans deploying LLMs.

Appendix

A A sample of the mock privacy policy

ORDEROO INC. PRIVACY POLICY

[...].

When you use our services, we process the following categories of personal data:

1. Your email address. You provide us with your email address when registering for the service. We use your email address for the following purposes: unique identifier, it serves as a unique identifier allowing you to set up and log in to your account (contractual necessity); account access, to let you reset your password if you forget it (contractual necessity); transaction-related-communication, to send you receipts of your orders (legal obligation: to issue receipts, according to the Receipts Act); distribution of own advertising, to send you advertisements of our own services, new functionalities or new order options (legitimate interest: informing the consumers about the available offers and features, and promoting them); distribution of third-party marketing, to send you advertisements of vendors selling their products on our site (legitimate interest: to subsidize the price of the service with payments from the vendors we promote); tracking transaction history, we keep it as a part of your order history in case it becomes necessary to reveal it to investigative authorities (legal obligation: Accounting Act and Code of Criminal Procedure); we use the domain name part of your email when profiling (see the separate section at the bottom of the document). We share your email address with Cloud711 (processor), for the purpose of data storage and backup, i.e., storing our IT operations on their servers (legitimate interest: lowering the cost of operation and keeping the data safe); Microsoft (processor) for the purpose of facilitating communication, i.e., sending our own emails (legitimate interest in outsourcing the operation of email servers and protocols); CoolAccountants (processor) for the purpose of accounting, i.e., reviewing our financial records and keeping them in order (legal obligation: Accounting Act); FraudDetectors (processor) for the purpose of fraud detection (legitimate interest: not becoming a victim of fraud). We store your email for as long as you're using our services, i.e., until you delete your account, PLUS THREE MONTHS.

[...].

13. Your geolocation. We automatically collect your precise geolocation from your device using GPS. We use your geolocation for the following purposes: delivery address suggestion, we suggest an address we believe you are at to streamline you making the order (consent); vendor suggestion, we suggest you ordering from vendors located close enough for our delivery persons to be able to bring you the order (contractual necessity); delivery, we track your location to see where exactly you find yourself, to facilitate the order delivery (legitimate interest: to increase efficiency and accuracy of our delivery persons); clients' location analytics, we use your geolocation to keep track how many of our clients are located in a given town, district, or street, to be able to have a sufficient number of delivery persons working that area (contractual necessity); anonymization, we anonymize information about your geolocation to strip it of any data that could link it to you, including by removing precise-location and information, to turn it into anonymous data which we use further for analytics and purposes we don't have to disclose (legitimate purpose: to derive value from data that is no longer personal); to profile you (see the separate section at the bottom of the document), to resolve issues related to deliveries or missing orders (legitimate interest: to improve the quality of our services). We share your geolocation with our delivery persons (processor) for the purpose of order delivery, to help them find you (contractual necessity); we share your geolocation with vendors close to you (controller), to facilitate their own marketing (legitimate

interest: building the relationship with vendors); we share your geolocation with Facebook (controller), for the purpose of ad delivery, i.e. verifying the efficacy of their ads and sending you third party marketing (legitimate interest: to make profit of available data) and Google (controller), for the purpose of ad delivery, i.e. verifying the efficacy of their ads and sending you third party marketing (legitimate interest: to make profit of available data). We store your geolocation info until you delete your account, PLUS THREE MONTHS.

Funding Przemysław Pałka: The research leading to these results has received funding from the Norwegian Financial Mechanism 2014–2021, project no. 2020/37/K/HSS/02769, titled “Private Law of Data: Concepts, Practices, Principles & Politics.” Pałka’s work between September 3 and October 3, 2023, was funded by and conducted within the frame of the Visiting Researcher program at the Department of Legal Studies, University of Bologna, Italy. Giovanni Sartor has been supported by the ERC Advanced Grant “CompuLaw” under the EU Horizon 2020 programme (Grant Agreement n. 833647). This work was partially supported by the following projects: CompuLaw – Computable Law – funded by the ERC under the Horizon 2020 (Grant Agreement N. 833647); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); PRIN2022 EQUAL – EQUitable Algorithms (Ref. Prot n. 2022KFLF3E\$_001 - CUP J53D23005560001); CLAUDETTE IV, founded by the EUI Research Council for founding; “FAIR - Future Artificial Intelligence Research” -- Spoke 8 “Pervasive AI,” under the European Commission’s NextGeneration EU programme, PNRR -- M4C2 -- Investimento 1.3, Partenariato Esteso (PE00000013). Marco Lippi was supported by CAI4DSA actions (Collaborative Explainable neuro-symbolic AI for Decision Support Assistant), PARTENARIATO ESTESO “Future Artificial Intelligence Research - FAIR,” SPOKE 1 “Human-Centered AI” Universi di Pisa, CUP B13C23005640006.” The proofreading and Open Access of the publication has been supported by a grant from the Faculty of Law and Administration under the Strategic Programme Excellence Initiative at the Jagiellonian University, Krakow, Poland.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Azamfirei R, Kudchadkar SR, Fackler J (2023) Large language models and the perils of their hallucinations. *Crit Care* 27(1):1–2
- Bakos Y, Marotta-Wurgler F, Trossen DR (2014) Does anyone read the fine print? consumer attention to standard-form contracts. *J Leg Stud* 43(1):1–35
- Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, Lovenia H, Ji Z, Yu T, Chung W, et al (2023) A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. <https://arxiv.org/abs/2302.04023>
- Bar-Gill O, Ben-Shahar O, Marotta-Wurgler F (2017) Searching for the common law: The quantitative approach of the restatement of consumer contracts. *The University of Chicago Law Review* pp 7–35
- Baron JR, Rollings NW, Oard DW (2023) Using chatgpt for the foia exemption 5 deliberative process privilege. In: *LegalAIIA@ ICAIL*, pp 32–48

- Ben-Shahar O, Chilton A (2016) Simplification of privacy disclosures: an experimental test. *J Legal Stud* 45(S2):S41–S67
- Calo R (2011) Against notice skepticism in privacy (and elsewhere). *Notre Dame L Rev* 87:1027
- Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz DM, Aletas N (2021) LexGLUE: A benchmark dataset for legal language understanding in english. <https://arxiv.org/abs/2110.00976>
- Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y et al (2023) A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*. <https://doi.org/10.1145/3641289>
- Citron DK (2016) The privacy policymaking of State attorneys general. *Notre Dame L Rev* 92:747
- Contissa G, Docter K, Lagioia F, Lippi M, Micklitz HW, Palka P, Sartor G, Torroni P et al (2018) Automated processing of privacy policies under the EU general data protection regulation. *Front Artif Intell Appl* 313:51–60
- Costante E, Sun Y, Petković M, Den Hartog J (2012) A machine learning solution to assess privacy policy completeness: (short paper). In: *Proceedings of the 2012 ACM workshop on privacy in the electronic society*, pp 91–96
- Davis KE, Marotta-Wurgler F (2019) Contracting for personal data. *NYUL Rev* 94:662
- DellaPosta D, Shi Y, Macy M (2015) Why do liberals drink lattes? *Am J Sociol* 120(5):1473–1511
- Emami-Naeini P, Dheenadhayalan J, Agarwal Y, Cranor LF (2021) An informative security and privacy “nutrition” label for internet of things devices. *IEEE Secur Priv* 20(2):31–39
- Encarnacion E (2019) Boilerplate indignity. *Ind LJ* 94:1305
- F, Russell NC, Schaub F, et al (2013) The usable privacy policy project. In: *Technical report, Technical Report, CMU-ISR-13-119*, Carnegie Mellon University
- Gentile G, Lynskey O (2022) Deficient by design? The transnational enforcement of the gdpr. *Int Compar Law Q* 71(4):799–830
- Gerber N, Gerber P, Volkamer M (2018) Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Comput Secur* 77:226–261
- Gharib M, Giorgini P, Mylopoulos J (2020) An ontology for privacy requirements via a systematic literature review. *Journal on Data Semantics* 9:123–149
- Gibson J (2017) Boilerplate’s false dichotomy. *Geo LJ* 106:249
- Harkous H, Fawaz K, Lebre R, Schaub F, Shin KG, Aberer K (2018) Polisis: Automated analysis and presentation of privacy policies using deep learning. In: *27th USENIX Security Symposium (USENIX Security 18)*, pp 531–548
- Harris KAGCDoJ (2014) Making your privacy practices public: Recommendations on developing a meaningful privacy policy. Accessible at: <https://oag.ca.gov/news/press-releases/attorney-general-kamala-d-harris-issues-guide-privacy-policies-and-do-not-track>
- Jablonska A, Tagiuri G (2023) Rescuing transparency in the digital economy: in search of a common notion in eu consumer and data protection law. *Yearbook of European Law* p yeard005
- Kaminski ME (2019) The right to explanation, explained. *Berk Technol Law J* 34(1):189–218
- Katz DM, Bommarito MJ, Gao S, Arredondo P (2023) GPT-4 passes the bar exam. Available at SSRN 4389233
- Kelley PG, Bresee J, Cranor LF, Reeder RW (2009) A “nutrition label” for privacy. In: *Proceedings of the 5th Symposium on Usable Privacy and Security*, pp 1–12
- Kenny C (2018) The equifax data breach and the resulting legal recourse. *Brook J Corp Fin & Com L* 13:215
- Kokolakis S (2017) Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. *Comput Secur* 64:122–134
- Kumaraguru P, Cranor L, Lobo J, Calo S (2007) A survey of privacy policy languages. In: *Workshop on Usable IT Security Management (USM 07): Proceedings of the 3rd Symposium on Usable Privacy and Security*, ACM
- Kvedar J (2019) Back to the grind: Rethinking grindr’s accountability for user content. *S Cal Interdisc LJ* 29:541
- Liepina R, Micklitz HW, et al (2019) GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism. In: *CEUR Workshop Proceedings, CEUR-WS, vol 2385*
- Lippi M, Palka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, Torroni P (2019) Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artif Intell Law* 27:117–139
- Lippi M, Contissa G, Jablonowska A, Lagioia F, Micklitz HW, Palka P, Sartor G, Torroni P (2020) The force awakens: artificial intelligence for consumer law. *J Artif Intell Res* 67:169–190
- Marotta-Wurgler F (2016) Self-regulation and competition in privacy policies. *J Leg Stud* 45(S2):S13–S39
- Marshall SR, Naumann LP (2018) What’s your favorite music? Music preferences cue racial identity. *J Res Pers* 76:74–91

- McDonald AM, Cranor LF (2008) The cost of reading privacy policies. *Isjlp* 4:543
- Newell BC, Purtova N, Moon YE, Paterson HJ III (2023) Regulating the data market: The material scope of american consumer data privacy law. *U Pa J Int'l L* 45:1055
- Obar JA, Oeldorf-Hirsch A (2020) The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Inf Commun Soc* 23(1):128–147
- Oltramari A, Piraviperumal D, Schaub F, Wilson S, Cherivirala S, Norton TB, Russell NC, Story P, Reidenberg J, Sadeh N (2018) Privonto: a semantic framework for the analysis of privacy policies. *Semantic Web* 9(2):185–203
- Palka P, Pałosz R, Wiśniewska K (2023) Annotated privacy policies of 100 online platforms. *Mendeley Data*. <https://doi.org/10.17632/pcgvm6zh43.1>
- Palmirani M, Martoni M, Rossi A, Bartolini C, Robaldo Livio (2018) Pronto: Privacy ontology for legal reasoning. In: Kō Andrea, Francesconi Enrico (eds) *Electronic Government and the Information Systems Perspective: 7th International Conference, EGOVIS 2018, Regensburg, Germany, September 3–5, 2018, Proceedings*. Springer International Publishing, Cham, pp 139–152. https://doi.org/10.1007/978-3-319-98349-3_11
- Pollach I (2011) Online privacy as a corporate social responsibility: an empirical study. *Bus Ethics A Eur Rev* 20(1):88–102
- Ravichander A, Black AW, Wilson S, Norton T, Sadeh N (2019) Question answering for privacy policies: Combining computational and legal perspectives. <https://arxiv.org/abs/1911.00841>
- Raza A, Asif M, Akram M (2023) Give your hunger a new option: Understanding consumers' continuous intention to use online food delivery apps using trust transfer theory. *Inter Natl J Consumer Stud* 47(2):474–495
- Reich N, Micklitz HW (2014) The court and sleeping beauty: the revival of the unfair contract terms directive (uctd). *Common Market Law Rev* 51(3):771
- Reidenberg JR, Breaux T, Cranor LF, French B, Grannis A, Graves JT, Liu F, McDonald A, Norton TB, Ram-anath R et al (2015) Disagreeable privacy policies: mismatches between meaning and users' understanding. *Berkeley Tech LJ* 30:39
- Richards N, Hartzog W (2018) The pathologies of digital consent. *Wash UL Rev* 96:1461 Sadeh N, Acquisti A, Breaux TD, Cranor LF, McDonald AM, Reidenberg JR, Smith NA, Liu
- Savelka J, Ashley KD, Gray MA, Westermann H, Xu H (2023) Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? arXiv: <https://arxiv.org/abs/2306.13906>
- Solove DJ (2021) The myth of the privacy paradox. *Geo Wash L Rev* 89:1
- Solove DJ, Hartzog W (2014) The fit and the new common law of privacy. *Colum L Rev* 114:583
- Solove DJ, Schwartz PM (2020) *Information privacy law*. Aspen Publishing
- Solow-Niederman A (2022) Information privacy and the inference economy. *Nw UL Rev* 117:357
- Sterkin SD (2004) Challenging adhesion contracts in california: a consumer's guide. *Golden Gate UL Rev* 34:285
- Tang C, Liu Z, Ma C, Wu Z, Li Y, Liu W, Zhu D, Li Q, Li X, Liu T, et al (2023) PolicyGPT: Automated analysis of privacy policies with large language models. arXiv preprint <https://arxiv.org/abs/2309.10238>
- Tøndel IA, Nyre ÅA (2012) Towards a similarity metric for comparing machine-readable privacy policies. In: *Open problems in network security: IFIP WG 11.4 International Workshop, iNetSec 2011, Lucerne, Switzerland, June 9, 2011, Revised Selected Papers*, Springer, pp 89–103
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Adv Neural Inf Process Syst*. 30
- Vivarelli A (2020) The crisis of the right to informational self-determination. *Italian LJ* 6:301
- Wagner W, Walker W (2019) Incomprehensible!: a study of how our legal system encourages incomprehensibility, why it matters, and what we can do about it. Cambridge University Press
- Waldman AE (2018) Privacy, notice, and design. *Stan Tech L Rev* 21:74
- Waldman AE (2020) Cognitive biases, dark patterns, and the 'privacy paradox. *Curr Opin Psychol* 31:105–109
- White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC (2023) A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint <https://arxiv.org/abs/2302.11382>
- Yerby J, Vaughn I (2022) Deliberately confusing language in terms of service and privacy policy agreements. *Issues Inf Syst*. 23(2)
- Yeung K, Bygrave LA (2022) Demystifying the modernized european data protection regime: Cross-disciplinary insights from legal and regulatory governance scholarship. *Regul Govern* 16(1):137–155
- Yu F, Quartey L, Schilder F (2023) Exploring the effectiveness of prompt engineering for legal reasoning tasks. *Find Assoc Comput Linguist ACL* 2023:13582–13596

Zaeem RN, German RL, Barber KS (2018) Privacycheck: automatic summarization of privacy policies using data mining. *ACM Trans Internet Technol* 18(4):1–18

Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How does NLP benefit legal system: a summary of legal artificial intelligence. arXiv preprint <https://arxiv.org/abs/2004.12158>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Przemysław Pałka¹ · Francesca Lagioia^{2,3} · Rūta Liepina³ · Marco Lippi⁴ · Giovanni Sartor^{2,3}

✉ Przemysław Pałka
przemyslaw1.palka@uj.edu.pl

Francesca Lagioia
francesca.lagioia@unibo.it

Rūta Liepina
ruta.liepina@unibo.it

Marco Lippi
marco.lippi@unifi.it

Giovanni Sartor
giovanni.sartor@unibo.it

¹ Faculty of Law and Administration, Jagiellonian University, ul. Bracka 12, 31-007 Krakow, Poland

² Department of Legal Studies, University of Bologna, Via Zamboni 27/29, 40126 Bologna, Italy

³ European University Institute, Via dei Roccettini 9, 50014 San Domenico di Fiesole, Italy

⁴ Department of Information Engineering, University of Florence, Via di Santa Marta 3, 50139 Florence, Italy