








Detecting Vague Clauses in Privacy Policies: The Analysis of Data Categories Using BERT Models and LLMs

Giulia GRUNDLER ^{a,1} , Rūta LIEPIŅA ^{a,1} , Mariaceleste MUSICCO ^{a,1} ,
 Francesca LAGIOIA ^{a,b,2} , Andrea GALASSI ^{c,2} ,
 Giovanni SARTOR ^{a,b}  and Paolo TORRONI ^c 

^a *CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy*

^b *European University Institute, Law Department, Italy*

^c *DISI, Alma-AI, University of Bologna, Italy*

Abstract. Despite some improvements in compliance metrics after the implementation of the European General Data Protection Regulation (GDPR), privacy policies have become longer and more ambiguous. They often fail to fully meet GDPR requirements, thus leaving users without a reliable way to understand how their data is processed. We present a novel corpus composed by 30 privacy policies of online platforms and a new set of annotation guidelines, to assess the level of comprehensiveness of information. We focus on the processed categories of data, classifying each clause either as fully informative or as insufficiently informative. In our experimental evaluation, we perform 6 different classification and detection tasks, comparing BERT models and generative Large Language Models.

Keywords. privacy policies, GDPR, NLP, Machine Learning, Large Language Models

1. Introduction

Our research is motivated by the growing information asymmetry between online service providers and consumers, exacerbated by the pervasive collection of personal data, processed through AI. Despite increased concerns over data practices, Data Protection Authorities (DPA) and consumers often fail to detect GDPR violations, since privacy policies (PPs) are characterised by vagueness, complexities and legal technicalities [1,2]. Their analysis is a highly complex and time-consuming task [3]. Studies, including Zaem et al. [4], reveal that, while the GDPR has improved certain compliance metrics, privacy policies have become longer and more ambiguous, further hindering user understanding. Our preliminary research [5] indicates that many policies fail to meet GDPR standards. Thus, users cannot understand how their personal data is processed and make informed decisions. In this paper, we show that AI can be used to effectively support users and DPA to analyse and legally assess privacy policies.

Most of the NLP studies dealing with policies focus on the detection and summarisation of information. They rely on different classic machine learning techniques, includ-

¹Equal contribution

²Corresponding authors: Francesca Lagioia: francesca.lagioia@eui.eu, Andrea Galassi: a.galassi@unibo.it

ing SVMs, Naive Bayes, Decision Trees, and Random Forests, as well as on transformers (e.g., BERT). As examples, consider the following. The PI-Extract tool [6] facilitates the detection of personal data collected and processing operations. The PrivacyGuide benchmarking tool [7] summarises and classifies privacy policies focusing on data practices and risk levels. Cejas et al. [8] propose an AI-assisted method that identifies the GDPR-relevant information and then evaluates such information against 23 completeness criteria, supported by a questionnaire compiled by users.

As shown, most approaches are aimed at detecting and retrieving information. Those also dealing with classification tasks, focus on individual clauses, thus failing to capture instances of non-compliance, which results from the interrelationships between different clauses. This critical issue also emerged in our previous experiments, in the context of the CLAUDETTE project [5,9]. Indeed, we were unable to successfully identify omissions and vague clauses in privacy policies.

To overcome this problem and detect whether a policy, as a whole, provides comprehensive information, we propose a new tagging methodology to specify relations between clauses. As for the experimental evaluation, we address six classification and detection tasks, using both fine-tuned transformers models (LEGAL-BERT, DistilRoBERTa) and generative LLMs (Gemini, Llama).

2. Corpus and Guidelines

Privacy policies can be deemed unlawful, under articles 13 and 14 GDPR if: (1) they omit the information required by the regulation, (2) allow for data processing beyond the prescribed limits, and (3) use unclear language. In our previous research [5], we designed a methodology that reflects the overall aims of the GDPR in regards to collection and processing of personal data. In particular, we developed a “golden standard” for privacy policies, which encompasses three key dimensions:

- A Comprehensiveness of information: whether a policy meets all the information requirements under art. 13 and 14 GDPR, or fails to do so, either by not providing required items of information, or by providing them insufficiently or vaguely;
- B Substantive compliance: whether a policy only permits personal data processing that comply with the GDPR;
- C Clarity of expression: whether a policy is written in clear and precise language.

Under each dimension, we identified the set of relevant categories of clauses and the criteria for their evaluation [5]. For the purpose of this research, we focus on the first dimension (comprehensiveness) and, in particular, on the categories of personal data concerned.

We trained our classifiers on a source corpus, containing 30 English privacy policies of online platforms, available on the providers’ websites and gradually downloaded and analysed starting from December 2023. Policies are often available in different versions for different jurisdictions, thus we selected the most recent versions available to European consumers/data subjects. The documents were selected among the services included in the CLAUDETTE preexisting data set, covering 142 English Terms of Services. In selecting them, we rely on the following criteria: a) coverage of different market sectors [10,11,12,13], (b) number of users and (c) the platform’s global relevance. Overall, the corpus contains 6,866 sentences, with an average of 229 sentences per policy. The tagging was done using Gloss, a tool developed by Savelka and Ashley [14].

Tag	Mandatory attributes			Mandatory sub-element	Mandatory attributes		Optional sub-elements	Mandatory attributes	
	Name	Values	Meaning		Name	Values		Name	Values
cat	Level	1	sufficiently informative	Category	ID	C_i	Specification	Ref	C_n
		2	insufficiently informative		Type	Open Closed		Type	Open Closed
					Kind	Geo Usage etc.	SubCategory	Ref	C_n
					Optional attribute			Type	Open Closed
					Ref	C_n		Kind	Geo Usage etc.

Table 1. Categories of personal data concerned annotation scheme.

2.1. Guidelines

As noted above, here we focus on the comprehensiveness dimension, with regard to the specification of data categories being processed.

The corpus annotation was done at sentence and sub-sentence levels. We assumed that each clause related to the topic (CAT) can be classified as: (1) *fully informative*, if the concerned categories of personal data are comprehensively specified and not vague (LEVEL="1"); or (2) *insufficiently informative* in all the other cases (LEVEL="2"). This assessment depends on a set of mandatory and optional sub-elements and attributes. Thus, we identified hierarchical levels for annotation, as detailed in Table 1.

In particular, we described each *genus* of data (CATEGORY) specified in the clause through the following mandatory attributes:

- A *unique identifier* (ID), whose value is a numeric string corresponding to the location of the (term denoting the) *genus* in the policy.
- *The kind of data* (KIND), i.e., the terms describing the *genus* of data and so identifying its content, e.g., geolocation, payment, usage, contact information.
- *The type of kind* (TYPE), i.e., the precision of terms describing the *genus*. In this regard we differentiated between:
 - Open terms, i.e., terms that vaguely abstract and create ambiguity as to the scope of the *genus* (TYPE="Open"). For instance, "usage information" is a broad concept, which may include a variety of data, such as the amount of time spent in an app or website, the goods users search for and their browsing behaviour.
 - Closed terms, i.e., concrete terms that clearly and unambiguously identify the data to be processed. For instance, "payment information" only refers to a well circumscribed set of information, e.g., full name, credit card number and expiration date, or bank account.
- *The link* (REF= C_n) between top-level categories (CATEGORY), to capture the contextual relationships between them.

The description of a category (e.g., "geolocation data") may also include two optional sub-elements: specification (e.g., "such as") and sub-category (e.g., "GPS information"). SPECIFICATION introduces a list of sub-categories of data, each denoting a *species* of the given *genus*. It has the following mandatory attributes.

Sufficiently Informative (Level="1")		Insufficiently Informative (Level="2")	
Rule 1	Category Type = "Closed" No Specification No Subcategories	Rule 1	Category Type = "Open" No Specification No Subcategories
Rule 2	Category Type = "Closed" Specification Type = "Closed"/"Open" SubCategory Type = "Closed"/"Open"	Rule 2	Category Type = "Open" Specification Type = "Open" Subcategory Type = "Closed"/"Open"
Rule 3	Category Type = "Open" Specification Type = "Closed" SubCategory Type = "Closed"	Rule 3	Category Type = "Open" Specification Type = "Closed" Subcategory Type = "Open"

Table 2. Criteria for assessing clauses on data processing.

- *The type*, distinguishing between (i) exhaustive ("Closed") and (ii) open-end lists ("Open"). It is usually signalled by the wording through which the list is introduced. For example, "it means", "namely", "limited to", refer to closed catalogues, while, "such as", "for example", "including", introduce open lists.
- *The link* (REF=C_n) between the specification (SPECIFICATION) and each genus (CATEGORY) it refers to (e.g. between "such as" and "geolocation data").

SUBCATEGORY denotes the *species* of the given CATEGORY. Each one is described through the mandatory attributes *kind* and *type* – in line with those of the CATEGORY– and *link* (REF) between the species (SUBCATEGORY) and genus (CATEGORY) it refers to (e.g., between "GPS information" and "geolocation data").

Based on these hierarchical levels of annotation, we defined a set of rules, detailed in Table 2, to assess whether a clause is sufficiently informative. In the following, we illustrate the application of each rule through a set of examples. For better understandability, tags are specified in XML.

Examples of sufficiently informative clauses

[Level 1, Rule 1]: `<cat Level="1">In specific cases, we may ask you to provide <Category ID="C1" Kind="Gov" Type="Closed">a copy of a document</Category> or <Category ID="C2" Kind="Gov" Type="Closed">government-issued ID</Category> to verify your identity, location, and/or account ownership.</cat>` [Blizzard, updated on 01/09/2023]

The clause above is fully informative, even though the two categories of collected information (i.e., "copy of document" and "government-issued ID") are not further specified. Indeed, such categories may only include a limited sub-set of data items, such as name and surname, date and place of birth.

[Level 1, Rule 2]: `<cat Level="1"><Category ID="C4" Kind="Payment" Type="Closed">Data about your billing address</Category> and <Category ID="C5" Kind="Payment" Type="Closed">method of payment</Category>, <Specification Type="Open" Ref="C5"> such as <SubCategory Kind="Payment" Type="Closed" Ref="C5">bank details</SubCategory>, <SubCategory Kind="Payment" Type="Closed" Ref="C5">credit</SubCategory>, <SubCategory Kind="Payment" Type="Closed" Ref="C5">debit</SubCategory>, or <SubCategory Kind="Payment" Type="Closed" Ref="C5"> other payment card information </SubCategory> </Specification>.</cat>` [Apple, updated 22/12/2022]

This clause is sufficiently informative since, even if the specification introduces an open-ended list, the two main categories of data (i.e., “billing address” and “payment methods”), as well as the sub-categories specifying payment methods (i.e., “bank details”, “credit”, “debit”, “other payment card information”) point to unequivocal data.

[Level 1, Rule 3]: `<cat Level="1">We collect<Category ID="C7" Kind="UsageData" Type="Open">information about your interactions with the Airbnb Platform</Category>,<Specification Type="Closed" Ref="C7">namely<SubCategory Kind="Purchase" Type="Closed" Ref="C7">bookings you have made</Specification></SubCategory>.</cat>` [Airbnb, updated 25/01/2023]

This clause is sufficiently informative, even though the category “interactions with the platform” creates ambiguity about the *genus* of collected data. Indeed, the category is followed by a closed specification, containing one precise and unambiguous data item (*species*), i.e., the user’s bookings.

Examples of insufficiently informative clauses

[Level 2, Rule 1]: `<cat Level="2">We collect <Category ID="C7" Kind="Gen" Type="Open">your personal information</Category> in order to provide and continually improve our products and services.</cat>` [Amazon, updated 11/08/2023]

This clause clearly fails to be fully informative, since it only refers to “personal information”, which is indubitably vague and too broad to even qualify as a category of data.

[Level 2, Rule 2]: `<cat Level="2">We collect<Category ID="C10" Kind="UsageData" Type="Open">information about how you engage with the Platform</Category>, <Specification Type="Open" Ref="C10">including<SubCategory Kind="UsageData" Type="Open" Ref="C10"> information about the content you view</SubCategory>, <SubCategory Kind="UsageData" Type="Open" Ref="C10">the duration</SubCategory>and <SubCategory Kind="UsageData" Type="Open" Ref="C10">frequency of your use</SubCategory>, <SubCategory Kind="SocialInteraction" Type="Open" Ref="C10">your engagement with other users</SubCategory>, <SubCategory Kind="InternetHistory" Type="Closed" Ref="C10">your search history</SubCategory>and<SubCategory Kind="Settings" Type="Open" Ref="C10"> your settings</SubCategory></Specification>.</cat>` [TikTok, updated 19/11/2023]

This clause is insufficiently informative, since the users “engagement with the platform” remains abstract, as it is followed by an open ended list (“including”) of ambiguous information on how they interact with the service, i.e., “information about the content you view”, “your engagement with other users”, etc.

[Level 2, Rule 3]: `<cat Level="2"> <Category ID="C11" Kind="SocialInteraction" Type="Open">Information about your relationships and interactions</Category> between you, other people and organisations, <Specification Type="Closed" Ref="C11">namely <SubCategory Kind="SocialInteraction" Type="Open" Ref="C11">types of engagement</SubCategory>, and <SubCategory Kind="SocialInteraction" Type="Open" Ref="C11">other interactions</SubCategory> related to people and organisations</Specification>.</cat>`

This clause clearly fails to be fully informative, since the information about the user “relationships and interactions” is too broad and vague: even though the specification is closed (“namely”), the subcategories are still open (i.e. “types of engagement, events,

Element	#	Kind	#	Kind	#	Kind	#
documents	30	DeviceInfo	378	GeoInfo	105	InternetHistory	38
sentences	6866	Gen	337	Metadata	102	Images	37
cat	749	UsageData	257	UserProfileInfo	98	Financial	28
Sufficiently inf.	181	UserGenerated	196	CommunicationProv	78	Gov	26
Insufficiently inf.	568	BasicAccountInfo	184	SocialInteraction	71	AudioTyping	23
		ContactInfo	149	ContentPreferences	58	CriminalRecord	17
		Payment	137	Performance	52	Deidentified	8
		Purchase	137	Settings	52	LicensingInfo	7
		HealthFitness	129	IdentityVerificationInfo	46	ListFriendsInfo	6
		Demographic	113	ContactList	41	Languageanalysis	1

Element	Open	Closed	Tot
Category	499	174	673
SubCategory	988	1250	2238
Specification	451	7	458

Table 3. Composition of the final corpus.

and other interactions”). This example is merely hypothetical, since we could not retrieve such type of clauses in the dataset.

2.2. Annotation Process and Final Corpus

Annotation was performed independently by two legal experts. The guidelines were refined over 10 double-tagged documents, through an iterative process alternating independent annotation of documents, discussion of the disagreement and clarification of the guidelines.

We evaluated the inter-annotator agreement, calculating Cohen’s kappa [15] on the tagged elements at word-level. The presence of multiple overlapping tags results in a multi-label tagging of each word. Instead of computing a single aggregate agreement measure, we used three fine-grained ones:³ (i) for each sentence in the document, if it is CAT or not; (ii) for each term in a CAT sentence, if the term is CATEGORY, SUBCATEGORY, or neither; (iii) for each term in a CAT, if it is a SPECIFICATION or not.⁴ To compute the agreement, both annotators tagged documents that were randomly sampled among the 20 not used for the revision of the guidelines, amounting to about 9% of the sentences in the whole corpus. The scores are 0.97 and 0.96 for CAT and SPECIFICATION respectively, indicating almost perfect agreement, and 0.80 for CATEGORY/SUBCATEGORY, indicating strong agreement. The Cohen’s κ on the attributes are 0.94 for LEVEL, 0.71 for TYPE, and 0.72 for KIND, indicating almost perfect agreement for the first and good agreement for the remaining ones. The final corpus consists of 30 documents, 13 annotated by both experts and 17 annotated by either of the two. Other details are summarized in Table 3. The data set is publicly available.⁵

3. Experimental Setting

In this study we address six tasks:

- CAT classification: given a sentence, classify it as CAT or non-CAT.⁶
- LEVEL classification: given a CAT sentence, classify it as sufficiently informative (LEVEL=“1”) or insufficiently informative (LEVEL=“2”).

³CAT and SPECIFICATION tags cannot be confused with others, so it makes sense to measure the agreement on them independently. CATEGORY and SUBCATEGORY tags never overlap, but can sometimes be mixed.

⁴For the second and third measures, we considered the words that both annotators tagged as CAT, to avoid propagating potential errors.

⁵<https://github.com/nlp-unibo/Privacy-Policies-Compliance>

⁶We segmented the text in sentences using the spaCy library [16], modified to avoid segmenting at colons.

- **TYPE** classification: given a **CATEGORY** or **SUBCATEGORY**, classify it as Open or Closed.⁷
- **KIND** classification: given a **CATEGORY** or **SUBCATEGORY**, classify it as one of the 30 possible **KINDS**, the list of which can be found in Table 3.
- **CATEGORY/SUBCATEGORY** detection: given a **CAT** sentence, find the spans of text corresponding to **CATEGORIES** and **SUBCATEGORIES**.⁸
- **SPECIFICATION** detection: given a **CAT** sentence, find the spans of text corresponding to **SPECIFICATIONS**.

The detection tasks were evaluated in two settings: (i) the **BIO** tagging format, where each token/word is classified as **B-Class** (begin), **I-Class** (inside) or **O-Class** (outside) for each class in question, and (ii) the **IO** tagging format, where the **B** tokens are tagged as **I**, resulting in **I-Class** and **O-Class** only.

Experiments were conducted using train-validation-test splits with rate 60%-20%-20%, determined at the document level so that sentences of the same document belong to the same split. The splits were created manually to (a) balance their composition, especially with respect to the **LEVEL** and **KIND** attributes, and (b) include as many double-tagged documents as possible in the test and validation sets, to increase their quality.

For all tasks, we experimented with four models: two **BERT**-based models, **DistilRoBERTa** [17] and **LEGAL-BERT** [18], and two generative LLMs, **Gemini 1.5 Flash** [19] and **Meta Llama 3.1** [20].⁹ **LEGAL-BERT** and **DistilRoBERTa** were fine-tuned with a sequence classification head for the first four tasks (classifications) and a token classification head for the last two (detections). They were trained for 10 epochs in classification tasks and 20 epochs in detection tasks, with early stopping, a learning rate of $2e^{-5}$ and a batch size of 4 for **LEVEL** classification and 8 otherwise. For the generative models, prompt-tuning was conducted on Gemini, based on the results on the validation set. Prompts are mostly based on the definitions of the classes in the annotation guidelines.¹⁰ In few-shot mode, they contain all the examples of the training set (for **CAT** classification only the positive examples, i.e., the **CAT** tags). The same prompts were used for Llama, with minimal changes to accommodate the preferred prompting style of the model. For **CAT** classification and both detection tasks, the number of input examples in few-shot mode was reduced due to the more limited token context window of Llama compared to Gemini. For both generative models, the two detection tasks were designed as extractions of the significant portions of text, and the outputs were then converted to **BIO/IO** labels for comparison with the other models. For the **KIND** classification task, we experimented also with a variation of the zero-shot setting, which we name **zero_I-shot**, where we provide only the name of the classes, which we deem self-explainable, without providing their definition.

We want to stress that, while generative LLMs are powerful tools capable of generating sophisticated responses, comparing LLMs with traditional classification models is

⁷**SPECIFICATIONS** were excluded, despite having the **TYPE** attribute, because they have a completely different structure and because there are only 7 examples of Closed **SPECIFICATIONS** in the dataset.

⁸We designed the **CATEGORY/SUBCATEGORY** detection task with both classes together because they never overlap and can influence each other, while **SPECIFICATIONS** generally overlap with **SUBCATEGORIES**, so they were detected independently.

⁹We used the following implementation of the models: nlpaueb/legal-bert-base-uncased, distilbert/distilroberta-base, gemini-1.5-flash-001, meta-llama/Meta-Llama-3.1-8B-Instruct

¹⁰Prompts are available in our repository.

Model	Cat			Level			Type		
	cat	non-cat	Avg.	1	2	Avg.	Closed	Open	Avg.
Majority baseline	0.00	0.92	0.46	0.00	0.86	0.43	0.00	0.72	0.36
Random baseline	0.24	0.64	0.44	0.33	0.63	0.48	0.45	0.53	0.49
LEGAL-BERT	0.75	0.96	0.86	0.77	0.92	0.84	0.80	0.81	0.81
DistilRoBERTa	0.77	0.96	0.87	0.65	0.89	0.77	0.82	0.82	0.82
Gemini zero-shot	0.58	0.90	0.74	0.40	0.63	0.51	0.72	0.69	0.70
Gemini few-shot	0.72	0.95	0.84	0.58	0.86	0.72	0.82	0.81	0.82
Llama zero-shot	0.40	0.66	0.53	0.38	0.71	0.54	0.68	0.32	0.50
Llama few-shot	0.43	0.70	0.57	0.44	0.81	0.62	0.70	0.56	0.63

Table 4. Results for the CAT, LEVEL and TYPE classifications. We report the F1 score for the classes, along with their macro average.

non-trivial, as their performance can vary based on how prompts are phrased and their probabilistic nature. While there is a research effort towards the definition and categorization of standards for prompts [21], even small changes in the prompt can lead to different results, and running the same query multiple times might produce different outputs [22,23,24,25,26].

4. Results and Discussion

For each task, we report the F1 score obtained by each classifier for each class, as well as their macro-average. We also report the performance of two baselines: a classifier that outputs a random class and one that always predicts the majority class.¹¹

CAT classification: As detailed in Table 4, DistilRoBERTa obtains the best result, very similar to LEGAL-BERT and Gemini few-shot. All three reach an almost perfect score on the non-CAT majority class, and a score above 0.70 for the CAT class. Gemini zero-shot gets a good macro F1 score of 0.74, while Llama remains below 0.60 in both zero and few-shot modes.

LEVEL classification: As shown in Table 4, the best model on both classes is LEGAL-BERT, followed by DistilRoBERTa. As for CAT classification, the third best model is Gemini in few-shot mode, while Llama remains much lower.

TYPE classification: Table 4 shows that the best result is obtained by DistilRoBERTa for both Open and Closed classes, as well as their macro average. LEGAL-BERT and Gemini few-shot obtain almost the exact same score, while in zero-shot mode Gemini gets an average F1 of 0.70. Llama only reaches an average of 0.63 in few-shot and 0.50 in zero-shot mode.

KIND classification: As shown in Table 5, Gemini few-shot reaches the best macro F1 score, closely followed by both variations of its zero-shot mode. Llama is the second-best model. LEGAL-BERT and DistilRoBERTa only get to 0.42 and 0.40. These last two models have a good performance in identifying the more represented classes (such as DeviceInfo, Gen, GeoInfo, HealthFitness and UserGenerated), but completely fail at classifying the classes with just a few or no training samples. This shows how good the generative models are in classification tasks, with a lot of labels and possibly very few examples for some of them. It is interesting to notice that, Gemini is better than the fine-tuned models also in the first zero-shot task, with just the classes' names as information.

CATEGORY/SUBCATEGORY detection: Table 6 shows that the best model is DistilRoBERTa in both settings, closely followed by LEGAL-BERT. Gemini few-shot ob-

¹¹ Additional metrics such as precision and recall are reported in our repository.

	Kind	Maj	Rand	LB	DB	Gemini			Llama		
						zero _l	zero	few	zero _l	zero	few
AudioTyping	0.00	0.00	0.22	0.00	0.00	0.15	0.29	0.25	0.25	0.15	0.13
BasicAccountInfo	0.00	0.08	0.43	0.41	0.39	0.39	0.51	0.44	0.23	0.33	0.27
CommunicationProv	0.00	0.04	0.54	0.50	0.41	0.67	0.60	0.55	0.57	0.56	0.56
ContactInfo	0.00	0.00	0.56	0.54	0.45	0.59	0.48	0.45	0.46	0.47	0.47
ContactList	0.00	0.00	0.35	0.35	0.75	0.83	0.97	0.47	0.67	0.57	0.57
ContentPreferences	0.00	0.00	0.22	0.42	0.42	0.33	0.50	0.29	0.33	0.26	0.26
CriminalRecord	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deidentified	0.00	0.00	0.29	0.00	0.50	0.29	0.00	0.00	0.00	0.00	0.00
Demographic	0.00	0.00	0.58	0.55	0.63	0.54	0.66	0.46	0.46	0.48	0.48
DeviceInfo	0.00	0.10	0.80	0.78	0.83	0.88	0.88	0.76	0.74	0.77	0.77
Financial	0.00	0.00	0.40	0.13	0.54	0.65	0.80	0.57	0.57	0.59	0.59
Gen	0.17	0.02	0.71	0.72	0.00	0.76	0.74	0.35	0.47	0.37	0.37
GeoInfo	0.00	0.05	0.79	0.69	0.79	0.77	0.92	0.82	0.90	0.90	0.90
Gov	0.00	0.00	0.50	0.33	0.25	0.25	0.18	0.00	0.33	0.50	0.50
HealthFitness	0.00	0.04	0.69	0.75	0.81	0.87	0.86	0.90	0.87	0.89	0.89
IdentityVerificationInfo	0.00	0.00	0.19	0.10	0.20	0.43	0.35	0.11	0.29	0.11	0.11
Images	0.00	0.00	0.56	0.76	0.77	0.69	0.88	0.65	0.73	0.76	0.76
InternetHistory	0.00	0.00	0.00	0.00	0.56	0.80	0.84	0.42	0.37	0.43	0.43
Languageanalysis	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
LicensingInfo	0.00	0.00	0.00	0.00	1.00	0.67	0.67	0.00	1.00	1.00	1.00
ListFriendsInfo	0.00	0.00	0.00	0.00	0.67	0.67	1.00	0.00	1.00	1.00	1.00
Metadata	0.00	0.00	0.65	0.67	0.23	0.54	0.67	0.26	0.44	0.22	0.22
Payment	0.00	0.08	0.65	0.59	0.70	0.78	0.79	0.72	0.75	0.69	0.69
Performance	0.00	0.05	0.56	0.56	0.59	0.86	0.86	0.67	0.38	0.15	0.15
Purchase	0.00	0.03	0.64	0.58	0.68	0.69	0.69	0.57	0.47	0.50	0.50
Settings	0.00	0.00	0.72	0.55	0.74	0.74	0.83	0.67	0.69	0.64	0.64
SocialInteraction	0.00	0.05	0.17	0.35	0.21	0.30	0.29	0.20	0.00	0.18	0.18
UsageData	0.00	0.12	0.65	0.60	0.58	0.63	0.64	0.53	0.55	0.59	0.59
UserGenerated	0.00	0.02	0.60	0.59	0.62	0.54	0.76	0.52	0.58	0.57	0.57
UserProfileInfo	0.00	0.11	0.33	0.43	0.34	0.50	0.43	0.39	0.44	0.45	0.45
Avg	0.01	0.03	0.42	0.40	0.53	0.57	0.60	0.39	0.48	0.47	0.47

Table 5. Results for the KIND classification task. We report the F1 score for the classes, along with their macro average. In the Gemini and Llama columns, $zero_l$ refers to the zero-shot experiment with class labels only, while the prompt for $zero$ also contains a short description of each class.

tains the best scores for both B-CATEGORY and I-CATEGORY labels with BIO evaluation, while Llama and Gemini zero-shot obtain a score below 0.50 in both settings.

SPECIFICATION detection: Table 6 shows that the best model is LEGAL-BERT. It is closely followed by DistilRoBERTa, that is the best model at detecting the B-SPECIFICATION label. Gemini gets almost identical scores between zero and few-shot modes in IO setting, but it is totally unable to detect B-SPECIFICATION in zero-shot BIO setting. This shows that in zero-shot it is already good at detecting the list-like spans of text, but it does not detect the introductory word(s). The same consideration holds for Llama, with lower scores and with the difference that it does not learn to identify B-SPECIFICATION even in few-shot mode.

5. Conclusion

We presented a novel corpus comprising 30 privacy policies of online platforms from different market sectors. We defined a novel set of guidelines and manually annotated the policies, to assess the level of comprehensiveness of information, i.e., if a policy meets all the information requirements under Art. 13 and 14 GDPR. In particular, we focused on the clauses specifying categories of data to be processed, and we classified each clause

Model	Category-Subcategory						Specification			
	<i>B-C</i>	<i>B-S</i>	<i>I-C</i>	<i>I-S</i>	<i>O</i>	Avg.	<i>B-Sp</i>	<i>I-Sp</i>	<i>O</i>	Avg.
Majority baseline	0.00	0.00	0.00	0.00	0.69	0.14	0.00	0.00	0.74	0.25
Random baseline	0.07	0.13	0.10	0.20	0.29	0.16	0.04	0.35	0.41	0.27
LEGAL-BERT	0.60	0.80	0.53	0.73	0.86	0.70	0.70	0.92	0.93	0.85
DistilRoBERTa	0.63	0.81	0.53	0.74	0.88	0.72	0.73	0.87	0.90	0.84
Gemini zero-shot	0.35	0.54	0.21	0.42	0.76	0.46	0.00	0.73	0.87	0.53
Gemini few-shot	0.64	0.75	0.57	0.65	0.85	0.69	0.42	0.81	0.89	0.71
Llama zero-shot	0.40	0.40	0.26	0.32	0.75	0.43	0.02	0.43	0.78	0.41
Llama few-shot	0.29	0.34	0.37	0.30	0.76	0.41	0.17	0.62	0.82	0.54
			<i>I-C</i>	<i>I-S</i>	<i>O</i>	Avg.	<i>I-Sp</i>	<i>O</i>	Avg.	
Majority baseline			0.00	0.00	0.69	0.23	0.00	0.74	0.37	
Random baseline			0.19	0.35	0.41	0.32	0.44	0.53	0.48	
LEGAL-BERT			0.59	0.83	0.87	0.76	0.92	0.93	0.92	
DistilRoBERTa			0.62	0.81	0.87	0.77	0.89	0.92	0.90	
Gemini zero-shot			0.22	0.40	0.76	0.46	0.83	0.87	0.85	
Gemini few-shot			0.50	0.56	0.85	0.64	0.82	0.89	0.86	
Llama zero-shot			0.26	0.31	0.75	0.44	0.54	0.78	0.66	
Llama few-shot			0.30	0.28	0.76	0.45	0.65	0.82	0.73	

Table 6. Results for the detection tasks, both in BIO and IO formats. We report the F1 score for the classes, along with their macro average. In the name of the classes, we use C for CATEGORY, S for SUBCATEGORY and Sp for SPECIFICATION.

either as fully informative or as insufficiently informative. The analysis of the collected documents indicates the extent of the issue: 568 out of 749 relevant clauses (76%) were insufficiently informative. Moreover, no policy was fully compliant.

We engaged with four classification tasks and two detection tasks, using both fine-tuned transformer models and generative LLMs. The fine-tuned models (LEGAL-BERT and DistilRoBERTa) obtained the best results in almost all tasks, except for KIND classification. This seems to be due to the high number of classes (KIND) and very few samples for some of them, making it hard to accomplish the task. Here, the generative LLMs’ ability to learn with very few examples emerges: Gemini surpasses the fine-tuned models even in zero-shot mode, with just the classes’ names as task description.

In future work, we will expand the categories of clauses and consider their relationships, e.g., what data are processed for what purposes and what the related legal bases are. We also aim to increase the number of instances of the least represented classes. Concerning the experiments, we plan to create a pipeline that includes all tasks. Finally, we believe that our annotation schema can be easily applied to other languages, so we plan to extend our study as we have done for Terms of Services [27,28,29].

Acknowledgments

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.: 20224TP-EYC - CUP J53D23005130001); PRIN2022 EQUAL – EQUitableALgorithms (Ref. Prot n. 2022KFLF3E_001 - CUP J53D23005560001); CLAUDETTE IV, founded by the EU 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).

References

- [1] Milne GR, Culnan MJ, Greene H. A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing*. 2006;25(2):238-49.
- [2] Reidenberg JR, Breaux T, Cranor LF, French B, Grannis A, Graves JT, et al. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech LJ*. 2015;30:39.
- [3] McDonald AM, Cranor LF. The cost of reading privacy policies. *Isjlp*. 2008;4:543.
- [4] Zaeem RN, Barber KS. The effect of the GDPR on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*. 2020;12(1):1-20.
- [5] Contissa G, Docter K, Lagioia F, Lippi M, Micklitz HW, Pałka P, et al. Claudette meets GDPR: Automating the evaluation of privacy policies using artificial intelligence. *European Consumer Organisation (BEUC) Study Report*. 2018.
- [6] Bui D, Shin KG, Choi JM, Shin J. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*. 2021.
- [7] Tesfay WB, Hofmann P, Nakamura T, Kiyomoto S, Serna J. PrivacyGuide: towards an implementation of the EU GDPR on internet privacy policy evaluation. In: *Proceedings of the fourth ACM international workshop on security and privacy analytics*; 2018. p. 15-21.
- [8] Amaral O, Abualhaija S, Torre D, Sabetzadeh M, Briand LC. AI-enabled automation for completeness checking of privacy policies. *IEEE Transactions on Software Engineering*. 2021;48(11):4647-74.
- [9] Liepina R, Contissa G, Drazewski K, Lagioia F, Lippi M, Micklitz HW, et al. GDPR privacy policies in CLAUDETTE: challenges of omission, context and Multilingualism. In: *CEUR Workshop Proceedings*. vol. 2385. Sun SITE Central Europe/RWTH Aachen University; 2019. p. 1-7. Available from: <https://ceur-ws.org/Vol-2385/paper9.pdf>.
- [10] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*. 2019;27:117-39.
- [11] Ruggeri F, Lagioia F, Lippi M, Torroni P. Detecting and explaining unfairness in consumer contracts through memory networks. *Artificial Intelligence and Law*. 2022;30(1):59-92.
- [12] Jablonowska A, Lagioia F, Lippi M, Micklitz HW, Sartor G, Tagiuri G. Assessing the cross-market generalization capability of the claudette system. In: *Proc. JURIX 2021*. IOS Press; 2021. p. 62-7.
- [13] Lagioia F, Jablonowska A, Liepina R, Drazewski K. AI in search of unfairness in consumer contracts: the terms of service landscape. *Journal of Consumer Policy*. 2022;45(3):481-536.
- [14] Savelka J, Ashley KD. Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. In: *JURIX*. vol. 313 of *Frontiers in Artificial Intelligence and Applications*. IOS Press; 2018. p. 111-20.
- [15] Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;20:37-46.
- [16] Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength Natural Language Processing in Python. 2020.
- [17] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*. 2019;abs/1910.01108.
- [18] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androusoopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics; 2020. p. 2898-904.
- [19] Team G, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, et al.. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context; 2024. Available from: <https://arxiv.org/abs/2403.05530>.
- [20] Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al.. The Llama 3 Herd of Models; 2024. Available from: <https://arxiv.org/abs/2407.21783>.
- [21] White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *CoRR*. 2023;abs/2302.11382.
- [22] Gan C, Mori T. Sensitivity and Robustness of Large Language Models to Prompt Template in Japanese Text Classification Tasks. In: *PACLIC. Association for Computational Linguistics*; 2023. p. 1-11.
- [23] Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In: *ACL (1). Association for Computational Linguistics*; 2022. p. 8086-98.
- [24] Martínez E. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*. 2024.
- [25] Salinas A, Morstatter F. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks

- Affect Large Language Model Performance. In: Ku LW, Martins A, Srikumar V, editors. Findings of ACL. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics; 2024. p. 4629-51. Available from: <https://aclanthology.org/2024.findings-acl.275>.
- [26] Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In: ICLR; 2024. .
- [27] Galassi A, Drazewski K, Lippi M, Torroni P. Cross-lingual Annotation Projection in Legal Texts. In: COLING. International Committee on Computational Linguistics; 2020. p. 915-26. Available from: <https://doi.org/10.18653/v1/2020.coling-main.79>.
- [28] Drazewski K, Galassi A, Jablonowska A, Lagioia F, Lippi M, Micklitz H, et al. A Corpus for Multilingual Analysis of Online Terms of Service. In: NLLP@EMNLP 2021. Association for Computational Linguistics; 2021. p. 1-8. Available from: <https://aclanthology.org/2021.nllp-1.1>.
- [29] Galassi A, Lagioia F, Jablonowska A, Lippi M. Unfair clause detection in terms of service across multiple languages. *Artificial Intelligence and Law*. 2024:1-49. Available from: <https://doi.org/10.1007/s10506-024-09398-7>.