# Detecting Vague Clauses in Italian Privacy Policies using Transformers, LLMs, and Cross-Lingual Techniques

**Giulia Grundler[a,1], Mariaceleste Musicco[a,1], Andrea Galassi[b,*], Francesca Lagioia[a,c,**], Rūta Liepiņa[a], Giorgio Resta[d], Sara Roccu[d], Giovanni Sartor[a,c] and Paolo Torroni[b]**

[a]CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy
[b]DISI, University of Bologna, Italy
[c]European University Institute, Law Department, Italy
[d]Department of Law, University of Roma Tre, Italy
ORCID (Andrea Galassi): https://orcid.org/0000-0001-9711-7042

**Abstract.** Privacy policies often fall short of providing a comprehensive account of how personal data is used, thus failing to comply with GDPR requirements. By doing so, they hamper the users' ability to make informed decisions about using services while ensuring that their data is used properly and fairly. This calls for automatic tools that can effectively identify potentially unlawful policies. Some initial tools have been developed for policies written in English. However, in the EU, whose 24 spoken languages are an integral part of its cultural heritage, such tools must address multilingualism. To this end, we present a new corpus of Italian privacy policies, with clauses labelled by experts in data protection law, to indicate the level of comprehensiveness of information. We focus on the categories of data processed, classifying each clause as either sufficiently or insufficiently informative ("vague"). We perform 6 different classification and detection tasks, comparing the performance of BERT-based models and generative Large Language Models. We also perform cross-language experiments to evaluate whether a pre-existing English corpus or classifiers can be leveraged for Italian and, vice versa, whether our corpus is informative enough to generalize to other languages.

## 1 Introduction

In the European Union (EU), Privacy policies are governed by the General Data Protection Regulation (GDPR),[2] which seeks to ensure the lawful, fair, and transparent processing of personal data. GDPR article 12 requires such policies to be "concise, transparent, intelligible, and easily accessible," and formulated in "clear and plain language." These requirements implicitly demand that information be made available in the native languages of the intended audiences. However, despite some improvements since the GDPR's enactment, empirical studies show that privacy policies are not only growing longer and vaguer [46], but also continue to fall short of GDPR requirements [10, 29, 30, 32]. A recent study [15] found that none of the English-language privacy policies examined were fully GDPR compliant. This widespread lack of compliance highlights the urgency of developing effective compliance-monitoring tools.

There are various reasons why a privacy policy may fail to comply with GDPR. Under Articles 13 and 14, privacy policies can be deemed unlawful if they: (1) omit information required by the regulation, (2) allow data processing beyond GDPR limits, or (3) use unclear language. We focus on the first aspect and address the task of identifying, within a given privacy policy, the clauses that can be deemed unlawful because they do not fully provide all the information required by law. We refer to them as *insufficiently informative* or *vague*. We focus on clauses that describe the kinds of data to be processed through vague and open-ended definitions. Insufficiently informative clauses typically contain phrases like "we collect data about your use of our service" or open-ended expressions like "such as", "including", "for example". Such imprecise or deliberately vague descriptions make it difficult for consumers to understand what information is collected about them and how companies use it, undermining their ability to make informed decisions about whether and how to use services.

Previous work on this task has focused on English documents [15]. However, linguistic diversity is a foundational principle of the EU, standing as a key symbol of European historical, political, social and cultural diversity.[3] From a legal perspective, the EU's commitment to multilingualism plays a crucial role in ensuring legal certainty, clarity, transparency, and democratic accountability. Reflecting this, EU legislation is published in all official languages. [4]

In Europe, multilingualism plays an important role not only in public documents, but also in legal information provided by private actors, such as contracts and privacy policies. Thus, a multilingual approach facilitates the monitoring of market and personal data practices, since in the EU, consumer and data protection authorities, as well as non-governmental organisations, tend to operate in their languages.

* Corresponding Author. Email: a.galassi@unibo.it.
** Corresponding Author. Email: francesca.lagioia@eui.eu.
[1] Equal contribution
[2] https://gdpr.eu/.
[3] See the Communication of 22 November 2005 "A New Framework Strategy for Multilingualism" (COM(2005) 596 final), available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum:c11084
[4] Judgment of the Court of 6 October 1982, C-283/81 - CILFIT, ECLI:EU:C:1982:335, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:61981CJ0283, paragraph 18.

Despite technological advances, such as Large Language Models (LLMs), advancements in multilingual legal applications are delayed due to the nature of legal language, which does not merely vary lexically across different languages, but also reflects distinct cultures and conceptual frameworks. A single term may carry divergent legal meanings depending on the jurisdiction, institutional context, or interpretive tradition in which it is embedded. This semantic and normative variability complicates cross-linguistic alignment, making it difficult for multilingual models to achieve reliable performance in tasks such as clause classification or compliance assessment. These challenges are further compounded by the scarcity of annotated datasets in some European languages, especially those of less populated Member States.

In this work, we focus on the automatic analysis of Italian privacy policies to examine methodologies for developing robust multilingual legal tools. In particular, we investigate whether it is necessary to build language-specific models and corpora, or if multilingual techniques can support scalable solutions across languages. We address the following research questions:

1. When an annotated Italian corpus is available, what is the most effective method for detecting insufficiently informative clauses?
2. Can knowledge be transferred across languages, by leveraging existing English resources?
3. Is an Italian corpus informative enough to enable the task in other languages, e.g., English?

To this end, we build the first Italian corpus for insufficiently informative clause detection, consisting of 30 annotated privacy policies. We use it to perform an extensive experimentation with multiple classification tasks in detecting insufficiently informative clauses in Italian. Our analysis considers a variety of models and learning paradigms, for a total of 84 experiments. Then, we use Grundler et al. [15]'s English corpus and classifiers trained on it to perform the task in Italian. We experiment with several cross-language techniques, including machine translation and multilingual models, for a total of 32 experiments. Finally, we evaluate the effectiveness of our Italian corpus as a resource for training models to perform the task on English documents. We do that by way of cross-language experiments. To the best of our knowledge, this is the first time such a task is addressed by applying cross-language techniques.

Our results indicate that BERT-based models outperform LLMs, and that English and Italian training data may be used interchangeably with comparable results. This suggests that our corpus could be a valuable resource to train models for other languages.

We make our corpus and our code publicly available.[5]

## 2 Related Work

### 2.1 *Automatic Analysis of Privacy Policies.*

Clause classification and information extraction from privacy policies have been widely explored with machine learning, mainly to detect and summarise key information for consumer understanding.

Early tools such as PI-Extract [6] used a BLSTM-CRF model on a semi-automatically annotated corpus of 30 policies to identify types of personal data and associated actions (e.g., collection, sharing). PrivacyGuide [40] classified and summarised privacy policies by data practices and risk levels, using a manually annotated dataset of 45 documents and experimenting with traditional classifiers such as Naive Bayes and SVMs. Moving toward compliance assessment, Amaral et al. [2] developed an AI-assisted method to check GDPR-required elements (e.g., purposes, legal basis) against 23 completeness criteria supported by a user questionnaire, while AutoCompliance [24] employed BERT-based models to detect ten types of GDPR Article 13 information from 304 policies and flag compliance issues.

With the rise of large language models (LLMs), recent work has benchmarked their performance. PolicyGPT [39] outperforms traditional models in sentence classification, while Torrado et al. [43] show that GPT-3 and GPT-4 can reliably identify types of data collected, offering competitive alternatives to earlier approaches.

By contrast, relatively few studies target vagueness in privacy policies, and to our knowledge, none focus on Italian texts. Early work by Reidenberg et al. [33] analysed five privacy policies to develop a grounded theory of vague and ambiguous terms, introducing a manual scoring method to compare sectors and highlighting differences in U.S. legal frameworks. Lebanoff and Liu [19] built neural models for vague term detection using a manually annotated corpus, testing context-aware and adversarial approaches. Liu et al. [22] trained neural networks to generate embeddings capturing both semantic and vagueness-related features, later explored through a visualisation tool. Building on this, Lian et al. [20] proposed the F-vague-Detector, which identifies "supporting evidence" and distinguishes four patterns of false vagueness. Most recently, Malik et al. [27] enhanced vagueness annotations on 100 policies, using entropy and standard deviation measures, and pioneered transformer-based models for sentence-level vagueness scoring.

Finally, work has begun to address insufficiently informative clauses. Grundler et al. [15] created a novel dataset and benchmarked two monolingual BERT models alongside LLMs. Our contribution extends this line by testing multilingual models, advanced prompting strategies, and cross-lingual transfer learning, with a focus on Italian privacy policies - a setting largely unexplored to date.

### 2.2 *Multilingual Transfer Learning.*

Multilingualism in natural language processing can significantly reduce the need for language-specific annotation and model training, by means of cross-lingual transfer learning, which leverages existing resources in a different language. However, it is seldom applied in the legal domain because of the specific technical terms that are often difficult to translate, the scarcity of multilingual legal datasets, and because a legal document often pertains to a specific jurisdiction linked to a single specific language. Isbister et al. [17] explore the effectiveness of building monolingual models for low-resource languages compared to using machine translation with existing English models. Their case study on Scandinavian languages for sentiment analysis shows that machine translation often yields better results. Chalkidis et al. [8] investigate cross-lingual transfer in legal NLP using multilingual models, leveraging the MULTIEURLEX dataset with EU laws translated into 23 languages. They find that, while multilingual models perform slightly worse than monolingual ones in same-language settings, they suffer a significant drop in performance when fine-tuned and tested on different languages, though this can be partially improved with adaptation techniques. Work by Artetxe et al. [3] shows the potential of translating the test set into English and running inference with a monolingual model, by using a stronger machine translation system that addresses the mismatch between the original test and the machine-translated one. However, their findings show that the optimal approach is highly task-dependent. Galassi et al. [12] compare test-translation and train-translation with annotation projection, from the English version of

[5] https://github.com/nlp-unibo/Privacy-Policies-Compliance

**Table 1.** Composition of our Italian corpus and comparison with the English corpus by Grundler et al. [15].

| Element | Ita | Eng |
|---|---|---|
| documents | 30 | 30 |
| sentences | 5862 | 6156 |
| cat | 875 | 770 |
| *Sufficiently inf.* | 220 | 201 |
| *Insufficiently inf.* | 655 | 569 |

| Element | Open | | Closed | | Tot | |
|---|---|---|---|---|---|---|
| | Ita | Eng | Ita | Eng | Ita | Eng |
| Category | 535 | 504 | 202 | 211 | 737 | 715 |
| SubCategory | 1187 | 1004 | 1150 | 1226 | 2337 | 2230 |
| Specification | 475 | 454 | 3 | 8 | 478 | 462 |

| Kind | Ita | Eng | Kind | Ita | Eng | Kind | Ita | Eng |
|---|---|---|---|---|---|---|---|---|
| DeviceInfo | 428 | 376 | GeoInfo | 138 | 111 | Images | 27 | 36 |
| Gen | 427 | 338 | Metadata | 104 | 101 | InternetHistory | 17 | 33 |
| UsageData | 301 | 277 | UserProfileInfo | 97 | 97 | Financial | 18 | 27 |
| UserGenerated | 200 | 203 | CommunicationProv | 73 | 75 | Gov | 22 | 24 |
| BasicAccountInfo | 172 | 190 | SocialInteraction | 41 | 68 | AudioTyping | 10 | 21 |
| ContactInfo | 202 | 176 | ContentPreferences | 35 | 58 | CriminalRecord | 10 | 16 |
| Purchase | 132 | 146 | Settings | 36 | 56 | LicensingInfo | 6 | 6 |
| Payment | 167 | 137 | IdentityVerificationInfo | 55 | 47 | ListFriendsInfo | 6 | 6 |
| HealthFitness | 126 | 120 | Performance | 38 | 44 | Deidentified | 2 | 3 |
| Demographic | 161 | 113 | ContactList | 22 | 40 | Languageanalysis | 1 | 0 |

each document into the corresponding sentences of the document in the target language. Other approaches include the use of an adversarial training and a language arbitration framework to make the LM embeddings as language-invariant as possible [5], or learning an alignment between word embeddings in different languages [18, 45], and use this sort of mapping function to transfer features from one language into another.

## 3 Corpus and Guidelines

### *3.1 Data Source*

Our starting point is the corpus produced by Grundler et al. [15], consisting of 30 English privacy policies from online platforms, as detailed in Table 1. Its annotations concern comprehensiveness of information, with a focus on the categories of personal data. We collected the corresponding Italian privacy policies. For each policy, we used the version uploaded or modified by the platform on a date close to the upload or modification for the corresponding policy in English, to maximize the coherence between our corpus and the one in English.

A comparative analysis of our documents with the English version revealed notable differences. Italian policies are consistently longer, largely due to the relative verbosity of the language. However, some discrepancies go beyond mere linguistic expansion. For instance, Yahoo's Italian policy (7,472 words) is significantly longer than its English counterpart (2,785 words) and provides more detailed descriptions of data collected and processing contexts. Similar divergences were found for Uber (IT: 9,584; EN: 7,938), Blizzard (IT: 7,331; EN: 5,274), and PayPal (IT: 8,320; EN: 6,532).

### *3.2 Guidelines*

Based on the methodology developed by Grundler et al. [15], the corpus annotation was done at sentence and sub-sentence levels, focusing on the clauses informing the user of the personal data processed by the online service. We assumed that each clause on the categories of data concerned (CAT) can be classified as: (1) *sufficiently 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

**Table 2.** Criteria for assessing clause vagueness, from Grundler et al. [15]. "Any" refers to both "Open" and "Closed" values.

| | Sufficiently Informative Level="1" | | Insufficiently Informative Level="2" |
|---|---|---|---|
| Rule 1 | Category = "Closed"<br>No Specification<br>No Subcategories | Rule 1 | Category = "Open"<br>No Specification<br>No Subcategories |
| Rule 2 | Category = "Closed"<br>Specification = Any<br>SubCategory = Any | Rule 2 | Category = "Open"<br>Specification = "Open"<br>Subcategory = Any |
| Rule 3 | Category = "Open"<br>Specification = "Closed"<br>SubCategory = "Closed" | Rule 3 | Category = "Open"<br>Specification = "Closed"<br>Subcategory = "Open" |

a set of mandatory and optional sub-elements and attributes. In particular, we described each category of data (CATEGORY) specified in the clause through the following mandatory attributes:

- ID: a numeric string corresponding to the location of the (term denoting the) category in the policy.
- KIND: the terms describing the category of data and so identifying its content, e.g., geolocation, payment, usage, contact information.
- TYPE: the precision of terms describing the category. In this regard we differentiated between:
  - *Open* terms, i.e., terms that vaguely abstract and create ambiguity as to the scope of the category. 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.
- REF: the link 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 specialization of the given category. It has the following

<CAT LEVEL="1"> To create <CATEGORY ID="C1" KIND="BASICACCOUNTINFO" TYPE="CLOSED"> an account </CATEGORY> you need to provide data <SPECIFICATION TYPE="OPEN" REF="C1"> including <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> your name </SUBCATEGORY> , <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> email address </SUBCATEGORY> and/or <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> mobile number </SUBCATEGORY> , and <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> a password </SUBCATEGORY> </SPECIFICATION>. </CAT>

[LinkedIn, 11/08/2020] from Grundler et al. [15].

<CAT LEVEL="1"> Per creare <CATEGORY ID="C1" KIND="BASICACCOUNTINFO" TYPE="CLOSED"> un account </CATEGORY> deve fornire alcuni dati che <SPECIFICATION TYPE="OPEN" REF="C1"> includono <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> il Suo nome </SUBCATEGORY> , <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> indirizzo email </SUBCATEGORY> e/o <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> numero di cellulare </SUBCATEGORY> , e <SUBCATEGORY KIND="BASICACCOUNTINFO" TYPE="CLOSED" REF="C1"> una password </SUBCATEGORY> </SPECIFICATION>. </CAT>

[LinkedIn, 11/08/2020] from our novel Italian corpus.

<CAT LEVEL="2"> We receive <CATEGORY ID="C1" KIND="GEN" TYPE="OPEN"> information about you </CATEGORY> from merchants as well as payment and transaction fulfillment providers, <SPECIFICATION TYPE="OPEN" REF="C1"> such as <SUBCATEGORY KIND="PURCHASE" TYPE="CLOSED" REF="C1"> payment confirmation details </SUBCATEGORY> , and <SUBCATEGORY KIND="PURCHASE" TYPE="CLOSED" REF="C1"> information about the delivery of products you have purchased </SUBCATEGORY> through our shopping features </SPECIFICATION>. </CAT>

[TikTok, 19/11/2023] from Grundler et al. [15].

<CAT LEVEL="2"> Riceviamo <CATEGORY ID="C1" KIND="GEN" TYPE="OPEN"> informazioni su di voi </CATEGORY> da commercianti nonché da fornitori di servizi per transazioni di pagamento e di acquisto, <SPECIFICATION TYPE="OPEN" REF="C1"> come ad esempio <SUBCATEGORY KIND="PURCHASE" TYPE="CLOSED" REF="C1"> le informazioni di conferma del pagamento </SUBCATEGORY> , nonché <SUBCATEGORY KIND="PURCHASE" TYPE="CLOSED" REF="C1"> sulla consegna dei prodotti da voi acquistati </SUBCATEGORY> attraverso le nostre funzionalità di shopping </SPECIFICATION>. </CAT>

[TikTok, 19/11/2023] from our novel Italian corpus.

**Figure 1.** Examples of annotated clauses in English (left) and Italian (right).

mandatory attributes.

- TYPE: whether the specification is an exhaustive ("Closed") or an open-ended list ("Open"). That 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.
- REF: The link between the specification and each category it refers to (e.g. between "such as" and "geolocation data").

SUBCATEGORY denotes the specialization 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 SUBCATEGORY and CATEGORY it refers to (e.g., between "GPS information" and "geolocation data").

Based on these hierarchical levels of annotation, a set of rules to assess whether a clause is sufficiently informative have been designed, as detailed in Table 2. The classification into sufficiently informative (Level 1) or insufficiently informative (Level 2) depends on whether the terms used are precise or vague, and whether any vagueness is resolved through further details. Essentially, the rules check whether vague terms (if any) are adequately anchored by specific, concrete descriptions, ensuring users understand exactly what data is being collected. Examples can be found in Figure 1.

### 3.3 Annotation Process and Validation

The corpus was annotated using Gloss [36] by a legal expert who is an Italian native speaker and fluent in English. Even if the guidelines include clear definitions of the labels, as with any manual annotation task on legal data, there is a margin of error that can be expected due to the length of the documents and the fine-grained level of labels. Consequently, we used the English dataset by Grundler et al. [15] to conduct a preliminary annotation training phase, allowing the annotator to refine their understanding of the guidelines before proceeding with the creation of the novel Italian corpus. The expert iteratively annotated three English policies and compared their result with the dataset's golden labels. This procedure was repeated three times, using different policies each time.

As validation, we measure the inter-annotator agreement between the expert's label in the last annotation round and the golden standard. We calculate the Cohen's $\kappa$ [9] on the labeled elements at word-level, with the same methodology used by Grundler et al. [15]. The three fine-grained metrics are: (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. The scores are 0.84 for CAT and CATEGORY/SUBCATEGORY, and 0.91 for SPECIFICATION, indicating strong agreement. The Cohen's $\kappa$ on the attributes are 0.77 for LEVEL, 0.74 for TYPE, and 0.71 for KIND, indicating good agreement. This indicates that our annotator and the annotators of Grundler et al. [15] have a similar interpretation of the guidelines.

### 3.4 Corpus

Our final corpus consists of 875 CAT clauses out of 5,862 total sentences. 220 clauses are labeled as sufficiently informative, and 665 as not. Table 1 shows detailed statistics on our Italian corpus and a comparison with the English corpus. Our corpus contains a higher number of CATs and elements, probably due to the higher average word count per document. Counterintuitively, the total number of closed elements is lower than in English. The dataset is publicly available.

## 4 Method

### 4.1 Task Definition

In this study, we follow Grundler et al. [15] and address the following six tasks. Each task is performed independently from the others.

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

- TYPE classification: given a CATEGORY or SUBCATEGORY, classify it as Open or Closed.
- KIND classification: given a CATEGORY or SUBCATEGORY, classify it as one of the 30 possible KINDs.
- CATEGORY/SUBCATEGORY detection: given a CAT sentence, find the spans corresponding to CATEGORIes and SUBCATEGORIes.
- SPECIFICATION detection: given a CAT sentence, find the spans of text corresponding to SPECIFICATIONs.

For the first task, we segment the text into sentences using the SPACY library. SPECIFICATIONs are excluded from the TYPE classification task, despite having the TYPE attribute, because they have a completely different structure and because there are only 3 examples of Closed SPECIFICATIONS in the dataset.

In designing the CATEGORY/SUBCATEGORY detection task, we group the two concepts together because they never overlap and can influence each other, while SPECIFICATIONS generally overlap with SUBCATEGORIES, so they are detected independently.

For the detection tasks, we consider 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. IO tagging is easier to learn, especially since the B class is often under-represented. However, the BIO tagging is required when one needs to distinguish between adjacent entities, as in the case of SUBCATEGORIES. We have therefore decided to experiment with both settings for both detection tasks to have a more complete experimental analysis.

Experiments were conducted using train-validation-test splits with rate 60%-20%-20%, determined at the document level with the same distribution used by Grundler et al. [15].

## 4.2 Models

For all tasks, we experimented with four BERT models: three Italian models (**ITALIAN-LEGAL-BERT** [21], **Italian BERT** [37] and **UmBERTo** [31]) and one multilingual model (**BERT multilingual** [11]), and two generative LLMs (**Gemini** 2.0 Flash [14] and **Llama**-3.1-8B-Instruct [25]).[6] We also made a preliminary evaluation with LLamantino [4], an LLM developed for Italian, before deciding to exclude it from our full experimentation to focus our computational budget on the most promising models. Indeed, the model consistently selected the first label shown in the prompt and often generated random or irrelevant responses.

BERT models were fine-tuned thrice with different seeds, with a sequence classification head for the first 4 tasks (classification) and a token classification head for the last 2 (detection). They were trained for 15 epochs in classification tasks and 25 epochs in detection tasks, with early stopping, a learning rate of $2e^{-5}$ and a batch size of 4 for INFORMATIVENESS classification and 8 otherwise.

For the generative LLMs, prompts are translations of the English prompts used in Grundler et al. [15], mostly based on the definitions of the classes in the annotation guidelines.[7] We experiment with zero-shot and 2 variations of few-shot mode. In the *static* version, we randomly select the examples from the training set, balancing the number of examples per class. In the *dynamic* version, inspired by the findings of Liu et al. [23] and Alfieri et al. [1], we dynamically select the same number of examples as the most similar to the given query, using a k-Nearest Neighbors algorithm based on the Euclidean distance between embeddings. In both cases, we select 15 examples for most tasks, 30 for Kind classification. For both generative models, the 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.

[6] We used the following implementation of the models: dlicari/Italian-Legal-BERT, dbmdz/bert-base-italian-cased, Musixmatch/umberto-commoncrawl-cased-v1, google-bert/bert-base-multilingual-cased, gemini-2.0-flash-001, meta-llama/Meta-Llama-3.1-8B-Instruct

[7] Prompts are available in our repository.

## 4.3 Transfer Learning Across Languages

Data annotation is notoriously costly, especially if it involves very specific domain knowledge or low-resource languages. It is therefore important to develop methods that minimize or eliminate the need for language-specific annotation. Multilingual NLP offers a potential solution to this challenge by enabling cross-lingual transfer learning, which allows models and corpora in one language to be exploited to perform tasks in other languages.

### 4.3.1 Transfer from English to Italian

We take inspiration from Galassi et al. [12], who study how to exploit the knowledge of a machine-learning system for English to build systems for German, Italian, and Polish. In our work, we Investigate how to develop a system for the Italian language without training on our Italian corpus. In particular, we use the English corpus of Grundler et al. [15] as the training source and evaluate the models on our Italian test set. We apply the following techniques.

- *Training set translation*. We machine translate the English training and validation sets to Italian and train new models for Italian. In this case, the system could be trained without the time-consuming activity of annotating a novel corpus.
- *Multilingual model*. We train a multilingual model in English. This also avoids both the need to annotate a new corpus and the need to translate, but still requires fine-tuning a new model.
- *Test set translation*. We use the ML models already trained in English as in [15] and, at inference time, we translate the Italian test set to English. This approach does not need a novel corpus nor the training of new systems, but requires using machine translation for each inference.

In addition to the Italian and multilingual BERT models detailed in Section 4.2, we experiment with the following English models: **DistilRoBERTa** [35], **LEGAL-BERT** [7] and **DeBERTa** [16].[8]

The translated versions of the corpora are generated with Opus-MT [41, 42], an open-source toolkit part of the Helsinki-NLP suite.[9] For each classification task, we translate the involved sentences (or portions of sentences) independently of each other and assign the original labels to the translations. We do not apply the methodology to the two detection tasks, as the labels are applied at the word-level and partially depend on their order, while translation may change the order of words, add new ones, or omit others.

### 4.3.2 Transfer from Italian to English

To evaluate the quality and the potential impact of our work, we want to investigate whether our novel corpus can be considered a valuable

[8] We used the following implementation of the models: nlpaueb/legal-bert-base-uncased, distilbert/ distilroberta-base, microsoft/deberta-v3-base

[9] We use the following models: opus-mt-tc-big-it-en, opus-mt-tc-big-en-it

**Table 3.** Classification results on our Italian test set. For each task, we report the F1-score of the classes and their macro average. The scores are the average value obtained over three runs. For the macro F1-score we also report the standard deviation.

| Method and Model | cat | | | | Informativeness | | | | Type | | | | Kind | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | cat | non-cat | Avg. | $\sigma$ | suff. | insuff. | Avg. | $\sigma$ | open | closed | Avg. | $\sigma$ | Avg. | $\sigma$ |
| *No learning* | | | | | | | | | | | | | | |
| gemini zero-shot | 0.75 | 0.95 | 0.85 | - | 0.56 | 0.72 | 0.64 | - | 0.63 | 0.70 | 0.67 | - | **0.56** | - |
| llama zero-shot | 0.61 | 0.92 | 0.77 | - | 0.51 | 0.77 | 0.64 | - | 0.72 | 0.66 | 0.69 | - | 0.44 | - |
| *Learning on Italian data* | | | | | | | | | | | | | | |
| gemini static few-shot | 0.72 | 0.94 | 0.83 | - | 0.61 | 0.79 | 0.70 | - | 0.59 | 0.70 | 0.65 | - | 0.54 | - |
| gemini dynamic knn few-shot | **0.76** | 0.95 | **0.86** | - | 0.64 | 0.79 | 0.72 | - | 0.76 | 0.77 | 0.76 | - | 0.54 | - |
| llama static few-shot | 0.57 | 0.89 | 0.73 | - | 0.56 | 0.78 | 0.67 | - | 0.57 | 0.69 | 0.63 | - | 0.45 | - |
| llama dynamic knn few-shot | 0.64 | 0.91 | 0.77 | - | 0.62 | 0.79 | 0.71 | - | 0.71 | 0.74 | 0.73 | - | 0.51 | - |
| ITALIAN-LEGAL-BERT | 0.72 | 0.95 | 0.83 | 0.008 | 0.62 | 0.86 | 0.74 | 0.027 | 0.79 | 0.76 | 0.78 | 0.005 | 0.37 | 0.020 |
| Italian BERT | **0.76** | **0.96** | **0.86** | 0.011 | 0.69 | 0.88 | 0.79 | 0.016 | **0.82** | **0.78** | **0.80** | 0.014 | 0.44 | 0.019 |
| UmBERTo | **0.76** | **0.96** | **0.86** | 0.008 | 0.67 | **0.89** | 0.78 | 0.043 | 0.79 | 0.74 | 0.76 | 0.013 | 0.34 | 0.036 |
| BERT multilingual | 0.75 | **0.96** | 0.85 | 0.010 | 0.68 | 0.86 | 0.77 | 0.009 | 0.81 | 0.77 | 0.79 | 0.016 | 0.41 | 0.021 |
| *Multilingual model* | | | | | | | | | | | | | | |
| BERT multilingual | 0.41 | 0.94 | 0.68 | 0.015 | 0.45 | 0.85 | 0.65 | 0.109 | 0.75 | 0.72 | 0.73 | 0.011 | 0.30 | 0.024 |
| *Training set translation* | | | | | | | | | | | | | | |
| ITALIAN-LEGAL-BERT | 0.71 | 0.95 | 0.83 | 0.004 | 0.68 | 0.87 | 0.78 | 0.030 | 0.76 | 0.75 | 0.76 | 0.008 | 0.38 | 0.007 |
| Italian BERT | 0.73 | **0.96** | 0.85 | 0.005 | 0.68 | 0.88 | 0.78 | 0.056 | 0.76 | **0.78** | 0.77 | 0.020 | 0.45 | 0.026 |
| UmBERTo | 0.73 | **0.96** | 0.85 | 0.003 | **0.72** | 0.88 | **0.80** | 0.022 | 0.73 | 0.73 | 0.73 | 0.024 | 0.32 | 0.049 |
| BERT multilingual | 0.75 | **0.96** | 0.85 | 0.017 | 0.67 | 0.88 | 0.78 | 0.026 | 0.78 | 0.76 | 0.77 | 0.019 | 0.43 | 0.001 |
| *Test set translation* | | | | | | | | | | | | | | |
| LEGAL-BERT | 0.74 | **0.96** | 0.85 | 0.016 | 0.67 | 0.87 | 0.77 | 0.015 | 0.80 | **0.78** | 0.79 | 0.006 | 0.45 | 0.015 |
| DistilRoBERTa | 0.73 | **0.96** | 0.85 | 0.008 | 0.71 | 0.88 | **0.80** | 0.013 | 0.79 | 0.75 | 0.77 | 0.012 | 0.45 | 0.003 |
| DeBERTa | 0.74 | **0.96** | 0.85 | 0.017 | **0.72** | **0.89** | **0.80** | 0.005 | 0.80 | **0.78** | 0.79 | 0.003 | 0.43 | 0.009 |

**Table 4.** Classification results on the English test set by Grundler et al. [15].

| Method and Model | cat | | | | Informativeness | | | | Type | | | | Kind | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | cat | non-cat | Avg. | $\sigma$ | suff. | insuff. | Avg. | $\sigma$ | open | closed | Avg. | $\sigma$ | Avg. | $\sigma$ |
| *Learning on English data* | | | | | | | | | | | | | | |
| LEGAL-BERT | 0.75 | **0.96** | **0.86** | 0.004 | **0.77** | **0.92** | **0.84** | 0.029 | **0.81** | **0.79** | **0.80** | 0.009 | **0.46** | 0.017 |
| *Training set translation* | | | | | | | | | | | | | | |
| LEGAL-BERT | **0.76** | **0.96** | **0.86** | 0.014 | 0.69 | 0.91 | 0.80 | 0.025 | 0.80 | 0.76 | 0.78 | 0.016 | 0.36 | 0.017 |
| *Test set translation* | | | | | | | | | | | | | | |
| Italian BERT | 0.74 | **0.96** | 0.85 | 0.012 | 0.76 | **0.92** | **0.84** | 0.022 | 0.80 | 0.75 | 0.77 | 0.031 | 0.35 | 0.013 |

resource for other languages. We experiment with the same transfer learning techniques detailed in the previous Section, but in the opposite direction: we use our corpus as a training source and the English corpus [15] as the test benchmark.

- *Training set translation*. We train one English model (LEGAL-BERT) with the Italian corpus translated into English.
- *Test set translation*. We train an Italian model (Italian BERT) on our corpus and we translate the English test set to Italian.

As a reference, we also experiment with a LEGAL-BERT model trained and tested on the English corpus.

## 5 Experimental Results

For each task, we report the F1 score obtained by each classifier for each class, as well as their macro-average. For Kind classification, we report only the macro-averaged F1 score. To account for the inherent stochasticity of the training process, we fine-tuned all BERT models three times and report the average result and standard deviation.

**Learning and testing on Italian data.** Overall, Italian BERT can be considered the best model for our task. As shown in Table 3, Italian BERT and UmBERTo yield the best result for CAT classification. BERT multilingual, Gemini in zero-shot setting, and Gemini in dynamic few-shot setting obtain similar results. Notably, the static few-shot prompt performs worse than the zero-shot one for both LLMs. As for INFORMATIVENESS classification, Italian BERT is the best model, followed by UmBERTo, with scores close to 0.80. In contrast, the best generative model, Gemini used in few-shot mode, reaches only 0.70 and 0.72, respectively with static and dynamic prompts. Italian BERT reaches the top-score of 0.80 also for TYPE classification, followed by BERT multilingual and ITALIAN-LEGAL-BERT. As for Kind classification, similarly to the findings of Grundler et al. [15], Gemini is the best model, reaching a F1 score of 0.56. We hypothesize that this may be caused by the abundance of classes of the task and the scarcity of instances in the dataset to represent some of them. For each task and LLM, the dynamic version of the few-shot prompt reaches a better score than the static one, confirming the importance of choosing examples close to the query. In the detection tasks, detailed in Table 5, Italian BERT and UmBERTo are always the best models. SPECIFICATIONS are easily detectable by all fine-tuned models, while performances for the CATEGORY-SUBCATEGORY task remain lower, especially concerning the identification of CATEGORIES. Finally, IO-mode performs better than BIO-mode, but with minimal differences for CATEGORY-SUBCATEGORY. However, it is worthwhile noticing that the former leads to a more ambiguous result than the latter, since it does not permit to distinguish between adjacent SUBCATEGORies. We hypothesize that general-purpose models may perform better than legal ones since the terms to detect are not domain-specific.

**Table 5.** Results for the detection tasks, both in BIO and IO formats. We use C to indicate CATEGORY, S for SUBCATEGORY and Sp for SPECIFICATION.

| Model | Category-Subcategory | | | | | | | Specification | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *B-C* | *I-C* | *B-S* | *I-S* | *O* | Avg. | $\sigma$ | *B-Sp* | *I-Sp* | *O* | Avg. | $\sigma$ |
| *BIO tagging* | | | | | | | | | | | | |
| gemini zero-shot | 0.48 | 0.51 | 0.71 | 0.72 | 0.85 | 0.65 | - | 0.64 | 0.51 | 0.80 | 0.65 | - |
| gemini few-shot | 0.66 | 0.60 | 0.81 | 0.77 | 0.88 | 0.75 | - | 0.51 | 0.69 | 0.82 | 0.67 | - |
| llama zero-shot | 0.23 | 0.23 | 0.32 | 0.47 | 0.76 | 0.40 | - | 0.14 | 0.32 | 0.71 | 0.39 | - |
| llama few-shot | 0.56 | 0.51 | 0.67 | 0.62 | 0.81 | 0.63 | - | 0.29 | 0.33 | 0.62 | 0.41 | - |
| ITALIAN-LEGAL-BERT | 0.68 | 0.66 | 0.79 | 0.77 | 0.89 | 0.76 | 0.013 | 0.68 | 0.88 | 0.92 | 0.83 | 0.004 |
| Italian BERT | 0.74 | 0.68 | 0.81 | 0.80 | 0.90 | **0.79** | 0.004 | 0.74 | 0.91 | 0.93 | **0.86** | 0.008 |
| UmBERTo | 0.67 | 0.64 | 0.84 | 0.80 | 0.91 | 0.77 | 0.009 | 0.69 | 0.94 | 0.96 | **0.86** | 0.009 |
| BERT multilingual | 0.76 | 0.64 | 0.81 | 0.78 | 0.88 | 0.77 | 0.005 | 0.69 | 0.90 | 0.93 | 0.84 | 0.007 |
| *IO tagging* | | | | | | | | | | | | |
| gemini zero-shot | | 0.56 | | 0.76 | 0.85 | 0.72 | - | | 0.52 | 0.80 | 0.66 | - |
| gemini few-shot | | 0.63 | | 0.81 | 0.88 | 0.77 | - | | 0.70 | 0.82 | 0.76 | - |
| llama zero-shot | | 0.27 | | 0.53 | 0.76 | 0.52 | - | | 0.36 | 0.71 | 0.54 | - |
| llama few-shot | | 0.54 | | 0.65 | 0.81 | 0.67 | - | | 0.34 | 0.62 | 0.48 | - |
| ITALIAN-LEGAL-BERT | | 0.74 | | 0.81 | 0.88 | 0.78 | 0.005 | | 0.89 | 0.92 | 0.90 | 0.006 |
| Italian BERT | | 0.70 | | 0.83 | 0.90 | **0.81** | 0.006 | | 0.89 | 0.92 | 0.91 | 0.020 |
| UmBERTo | | 0.68 | | 0.84 | 0.90 | **0.81** | 0.001 | | 0.92 | 0.94 | **0.93** | 0.006 |
| BERT multilingual | | 0.68 | | 0.81 | 0.89 | 0.79 | 0.004 | | 0.90 | 0.92 | 0.91 | 0.008 |

**Transfer from English to Italian.** Cross-lingual experiments from English to Italian lead to similar results to those obtained in the previous setting. Specifically, training in English and translating the test set at inference time lead to the top score for INFORMATIVENESS classification with DeBERTa. This result is particularly important because such a setting is the most convenient one in terms of computational resources. The scores for the other tasks are also high, with KIND classification reaching the best score among the fine-tuned models (0.45). Similarly to the findings of Chalkidis et al. [8], we notice that, while multilingual BERT reaches similar results to monolingual models in a same-language setting, its performance noticeably degrades when trained and tested in different languages.

**Transfer from Italian to English.** Table 4 shows the results of our experiments using our Italian corpus as a training set for English. In 3 tasks out of 4, using our corpus for training yields similar results to those obtained by LEGAL-BERT trained on the English corpus. The only exception is KIND classification, for which the model trained on the original corpus obtains a F1 score of 0.46, 10 points higher than our methods. Our two approaches yield similar results across all tasks, and neither of them is consistently better than the other. This indicates that our corpus can be used to generalize to other languages.

## 6 Error Analysis

We focus our error analysis on using English data for training and Italian data for testing. In particular, we analyze the results of DeBERTa in the *test set translation* approach, as it offers the best trade-off between resource efficiency and performance. The model is trained on the English corpus and tested on Italian documents automatically translated into English at inference time.

**AG:** @Legal If I understand correctly, I think the current error analysis contains errors

- If I understand correctly, the focus of the TikTok error is the mistranslation. Which is due to the lack of context. So, we should say (1) mistranslation may lead to misclassification, (2) since for TYPE and KIND task the context is very limited, mistranslation is more probable, and consequently errors, (3) present the example.
- For MySugr, I think the translation is correct: "esaminiamo" becomes "review". The problem is the mismatch between the original Italian lexicon and the original English lexicon. Even if the general semantic of the words is similar, in this domain that little difference has a strong impact.
- For MySugr, why does it matter if we are using only one version of the two (if the labels are correct)?
- So, I feel like we have swapped the analysis: for tiktok is a matter of mistranslation, not of nuances of language, while for mysugr is a matter of differences in the original language, not a matter of translation.
- We need to discuss the Paypal example. I think it is even different from MySugr, since it is not a matter of domain-specific lexicon, but of saying different things.

We identified two primary sources of misclassification: (i) divergences in concepts and terminology, and (ii) inaccuracies in the machine translation of the Italian dataset. These issues are not mutually exclusive and may appear together.

### 6.1 Conceptual and Terminological Divergences

Misclassifications frequently arise from linguistic nuances that reflect divergent legal or contextual meanings across languages. Consider the following example:

> *Deduciamo le vostre generalità (quali la fascia d'età e il genere) nonché gli interessi sulla base delle informazioni di cui disponiamo riguardo a voi.*
>
> [TikTok, 19/11/2023, Original Italian]

> *We infer your attributes (such as age-range and gender) and interests based on the information we have about you.*
>
> [TikTok, 19/11/2023, Original English]

When performing the TYPE and KIND classification tasks, the Italian term "genere" was incorrectly translated as "genre" (in the artistic sense) instead of "gender". This mistranslation led to the misclassification of TYPE and KIND as "Closed" and "Settings", respectively. However, such translation errors occur only for TYPE and KIND classification, where translation is performed over the isolated SUBCATEGORY span rather than the full sentence. As a result, the translation model lacks sufficient contextual information to disambiguate the term. In contrast, for the INFORMATIVENESS task, translation is

applied at the sentence level, providing the model with broader contextual cues that allow correct disambiguation and prevent misclassification.

### 6.2 *Translation Inaccuracies*

In other cases, the translations distorted the legal semantics of the source text. Consider the following clause:

> *Esaminiamo anche il tuo indirizzo IP per valutare da quale paese o regione stai usando i nostri servizi e per fornirti le funzioni e le informazioni rilevanti nel tuo paese.*
> [My Sugr, 01/01/2022, Original Italian]

> *We also review your IP address to assess which country or region you are using our services from and to provide you with the relevant functions and information in your country.*
> [My Sugr, Italian Translated into English]

> *We also process your IP address to assess from which country or region you are using our services and to provide you with the features and information which is relevant in your country.*
> [My Sugr, 01/01/2022, Original English]

The translation of "esaminiamo" as "review" rather than "process" understates the scope and legal significance of the data handling activity. While the English original uses "process" – a term with clear implications under the GDPR – the translated version may mislead both users and classifiers into interpreting the IP address handling as a superficial check, rather than a possible profiling operation.

### 6.3 *OTHER EXAMPLE*

AG: NEW EXAMPLE? @Giulia, @Celeste, @Ruta confirm (or change) and discuss

> *Inferences drawn to create a profile about you that may reflect behavior patterns and personal preferences, such as gender, income, browsing and purchasing habits, and creditworthiness.*
> [PayPal, 1/11/2023, Original English]

> *PayPal may draw conclusions about your behaviour patterns, personal preferences, browsing and purchasing habits, and creditworthiness.*
> [Paypal, Italian Translated into English]

> *Clause from Italian Privacy Policy: "PayPal può trarre conclusioni in merito ai modelli di comportamento, alle preferenze personali, alle abitudini di navigazione e di acquisto e all'affidabilità creditizia dell'Utente.*
> [PayPal, ??/??/????, Original Italian]

## 7 Limitations

Concerning the creation of the Italian corpus, a limitation is the use of only one human annotator, trained on the English corpus. Hence, we do not have a measure of agreement on the Italian corpus. While we acknowledge that agreement measured on the English corpus is only a proxy for the agreement on the Italian corpus, we shall stress that our task requires annotators with very specific expertise, who are also fluent in both Italian and English, and they are rare.

Concerning our experiments with LLMs, we acknowledge that, although generative LLMs are capable of producing complex responses, comparing their performance with that of traditional classifiers is not straightforward. Their behavior can be influenced by prompt formulation and inherent probabilistic variability. While ongoing research is working toward establishing standards and best practices for prompt design [44], minor changes or even repeated runs of the same prompt may lead to different outputs and variations in results [13, 26, 28, 34, 38].

Finally, as mentioned in Section 4.3, we did not experiment with cross-lingual techniques for the detection tasks, due to the impossibility of preserving a faithful ground truth using machine translation.

## 8 Conclusion

In this paper, we release a novel dataset of 30 Italian privacy policies annotated for the detection of insufficiently informative clauses. In particular, we focus on assessing descriptions of the data categories processed. We experiment with several models to solve the task in Italian and investigate whether it is possible to leverage existing resources in English to automatically identify such clauses. Our experiments show that BERT-based models yield better results than LLMs and that English resources can be effectively used to train a system in Italian with comparable results. The *test set translation* appears to be the best technique, despite some translation errors that we discuss in our error analysis, as it does not require a novel corpus nor the training of a new system. Similarly, we demonstrate that our Italian corpus can be effectively used to perform the task in English. These results, and the broader potential of automated privacy policy analysis in Italian and other languages, hold the promise of empowering consumers, supporting enforcement agencies, and assisting companies seeking to comply with data protection rules.

In the future, we will enrich the dataset with additional labels related to purposes and legal basis of data processing. We also aim to evaluate whether our result can be generalized to other languages, such as underrepresented European languages, for which legal corpora and reliable machine translation tools are often lacking. Moreover, we plan to investigate transfer between languages other than English to assess if our method is effective for any pair of European languages. This would also allow us to evaluate whether English as a lingua franca is necessary in this domain and context. Finally, we aim to build a full pipeline that includes all our tasks in succession.

## Acknowledgements

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 – EQUitableALgorithms (Ref. Prot n. 2022KFLF3E_001 - CUP J53D23005560001); CLAUDETTE IV under the EUI Research Council for funding; "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] F. Alfieri, G. Grundler, F. Galloni, R. Liepiņa, F. Lagioia, A. Galassi, and P. Torroni. Dynamic demonstrations selection for few-shot legal

argument mining. In *Proceedings of the First Argument Mining and Empirical Legal Research Workshop (AMELR 2025)*, 2025.

[2] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, and L. C. Briand. Ai-enabled automation for completeness checking of privacy policies. *IEEE Transactions on Software Engineering*, 48(11):4647–4674, 2021.

[3] M. Artetxe, V. Goswami, S. Bhosale, A. Fan, and L. Zettlemoyer. Revisiting machine translation for cross-lingual classification. In *EMNLP*, pages 6489–6499, 2023.

[4] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, and G. Semeraro. Llamantino: Llama 2 models for effective text generation in italian language, 2023.

[5] M. A. Bornea, L. Pan, S. Rosenthal, R. Florian, and A. Sil. Multilingual transfer learning for QA using translation as data augmentation. In *AAAI*, pages 12583–12591. AAAI Press, 2021.

[6] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *EMNLP (Findings)*, pages 2898–2904. ACL, 2020.

[8] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *EMNLP*, pages 6974–6996. Association for Computational Linguistics, 2021.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[10] G. Contissa, K. Docter, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Pałka, G. Sartor, and P. Torroni. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. *Available at SSRN 3208596*, 2018.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[12] A. Galassi, F. Lagioia, A. Jabłonowska, and M. Lippi. Unfair clause detection in terms of service across multiple languages. *Artificial Intelligence and Law*, pages 1–49, 2024. doi: 10.1007/s10506-024-09398-7.

[13] C. Gan and T. Mori. Sensitivity and robustness of large language models to prompt template in japanese text classification tasks. In *PACLIC*, pages 1–11. Association for Computational Linguistics, 2023.

[14] GeminiTeam. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

[15] G. Grundler, R. Liepina, M. Musicco, F. Lagioia, A. Galassi, G. Sartor, and P. Torroni. Detecting vague clauses in privacy policies: The analysis of data categories using BERT models and LLMs. In *JURIX*, volume 395, pages 72–83. IOS Press, 2024. doi: 10.3233/FAIA241235.

[16] P. He, J. Gao, and W. Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *ICLR*. OpenReview.net, 2023.

[17] T. Isbister, F. Carlsson, and M. Sahlgren. Should we stop training more monolingual models, and simply use machine translation instead? In *NoDaLiDa*, pages 385–390. Linköping University Electronic Press, Sweden, 2021.

[18] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. In *ICLR*, 2018.

[19] L. Lebanoff and F. Liu. Automatic detection of vague words and sentences in privacy policies. In *EMNLP*, pages 3508–3517. Association for Computational Linguistics, 2018.

[20] X. Lian, D. Huang, X. Li, Z. Zhao, Z. Fan, and M. Li. Really vague? automatically identify the potential false vagueness within the context of documents. *Mathematics*, 11(10):2334, 2023.

[21] D. Licari and G. Comandè. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *EKAW (Companion)*, volume 3256 of *CEUR Workshop Proceedings*, 2022.

[22] F. Liu, N. L. Fella, and K. Liao. Modeling language vagueness in privacy policies using deep neural networks. In *AAAI Fall Symposia*, 2016.

[23] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for gpt-3? In *DeeLIO@ACL*, pages 100–114, 2022.

[24] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang. Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13. In *Proceedings of the Web Conference 2021*, pages 2154–2164, 2021.

[25] LlamaTeam. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[26] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *ACL (1)*, pages 8086–8098. Association for Computational Linguistics, 2022.

[27] G. Malik, S. Yildirim, M. Cevik, and A. Bener. An empirical study on vagueness detection in privacy policy texts. In *Canadian AI*, 2023.

[28] E. Martínez. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, 2024.

[29] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[30] P. Pałka, F. Lagioia, R. Liepina, M. Lippi, and G. Sartor. Make privacy policies longer and appoint llm readers. *Artificial Intelligence and Law*, pages 1–33, 2025.

[31] L. Parisi, S. Francia, and P. Magnani. Umberto: an italian language model trained with whole word masking, 2020.

[32] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. McDonald, T. B. Norton, R. Ramanath, et al. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[33] J. R. Reidenberg, J. Bhatia, T. D. Breaux, and T. B. Norton. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190, 2016.

[34] A. Salinas and F. Morstatter. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. In *Findings of ACL*, pages 4629–4651. Association for Computational Linguistics, 2024.

[35] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, 2019.

[36] J. Savelka and K. D. Ashley. Segmenting U.S. court decisions into functional and issue specific parts. In *JURIX*, volume 313, pages 111–120. IOS Press, 2018.

[37] S. Schweter. Italian bert and electra models, 2020. URL https://doi.org/10.5281/zenodo.4263142.

[38] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *ICLR*, 2024.

[39] C. Tang, Z. Liu, C. Ma, Z. Wu, Y. Li, W. Liu, D. Zhu, Q. Li, X. Li, T. Liu, and L. Fan. Policygpt: Automated analysis of privacy policies with large language models. *CoRR*, abs/2309.10238, 2023.

[40] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna. 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*, pages 15–21, 2018.

[41] J. Tiedemann and S. Thottingal. OPUS-MT — Building open translation services for the World. In *EAMT*, 2020.

[42] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vázquez, and S. Virpioja. Democratizing neural machine translation with OPUS-MT. *Lang. Resour. Evaluation*, 58(2):713–755, 2024.

[43] D. R. Torrado, I. Yang, J. M. del Álamo, and N. Sadeh. Large language models: a new approach for privacy policy analysis at scale. *Computing*, 106(12):3879–3903, 2024.

[44] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382, 2023.

[45] R. Xu, Y. Yang, N. Otani, and Y. Wu. Unsupervised cross-lingual transfer of word embedding spaces. In *EMNLP*, pages 2465–2474, 2018.

[46] R. N. Zaeem and K. S. Barber. The effect of the gdpr on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):1–20, 2020.