Word-level Annotation of GDPR Transparency Compliance in Privacy Policies using Large Language Models

Thomas Cory
Technische Universität Berlin
Berlin, Germany
cory@tu-berlin.de

Wolf Rieder
Technische Universität Berlin
Berlin, Germany
w.rieder@tu-berlin.de

Julia Krämer
Erasmus University Rotterdam
Rotterdam, Netherlands
j.k.kramer@law.eur.nl

Philip Raschke
Technische Universität Berlin
Berlin, Germany
philip.raschke@tu-berlin.de

Patrick Herbke
Technische Universität Berlin
Berlin, Germany
p.herbke@tu-berlin.de

Axel Küpper
Technische Universität Berlin
Berlin, Germany
axel.kuepper@tu-berlin.de

Abstract

Ensuring transparency of data practices related to personal information is a core requirement of the General Data Protection Regulation (GDPR). However, large-scale compliance assessment remains challenging due to the complexity and diversity of privacy policy language. Manual audits are labour-intensive and inconsistent, while current automated methods often lack the granularity required to capture nuanced transparency disclosures.

In this paper, we present a modular large language model (LLM)-based pipeline for fine-grained word-level annotation of privacy policies with respect to GDPR transparency requirements. Our approach integrates LLM-driven annotation with passage-level classification, retrieval-augmented generation, and a self-correction mechanism to deliver scalable, context-aware annotations across 21 GDPR-derived transparency requirements. To support empirical evaluation, we compile a corpus of 703,791 English-language privacy policies and generate a ground-truth sample of 200 manually annotated policies based on a comprehensive, GDPR-aligned annotation scheme.

We propose a two-tiered evaluation methodology capturing both passage-level classification and span-level annotation quality and conduct a comparative analysis of seven state-of-the-art LLMs on two annotation schemes, including the widely used OPP-115 dataset. The results of our evaluation show that decomposing the annotation task and integrating targeted retrieval and classification components significantly improve annotation accuracy, particularly for well-structured requirements. Our work provides new empirical resources and methodological foundations for advancing automated transparency compliance assessment at scale.

Keywords

privacy, privacy policy, annotation, LLM, RAG, self-correction, GDPR

1 Introduction

The European Union's General Data Protection Regulation (GDPR), enacted in 2018, sets a global benchmark for data privacy and transparency. Among its key provisions, Articles 13 and 14 establish stringent requirements for organisations to provide clear and comprehensive information about data collection, processing, and storage practices [13].

Privacy policies, written in natural language, have emerged as the de facto standard for informing users about data practices and continue to persist as the primary means of communicating privacy-related information [29]. However, privacy policies are often long and complex [3], use ambiguous language [33], contain contradictions [31], or omit critical details [44]. These factors, coupled with the use of legal jargon, often make privacy policies difficult to understand for non-experts and hinder verification by regulators, making manual audits resource-intensive and prone to inconsistencies.

Prior research has highlighted the difficulty of manual compliance assessments and the need for scalable, automated solutions capable of systematically analysing privacy policies and identifying areas of potential non-compliance [42]. In response, studies have explored the use of machine learning (ML) [25] and natural language processing (NLP) techniques [9] to automate the analysis of privacy policies. However, existing approaches continue to face challenges in accurately capturing nuanced legal language and context, particularly regarding the annotation (i.e. identification and labelling) of relevant words and phrases within sentences.

In this work, we address these challenges by introducing a modular, large language model (LLM)-based pipeline for fine-grained, word-level annotation of privacy policies with respect to GDPR transparency requirements. Our approach leverages recent advances in LLMs, passage-level classification, and retrieval-augmented generation (RAG) to deliver context-aware, high-granularity annotations suitable for compliance analysis. The key contributions of this paper are as follows:

(1) We define an annotation scheme comprising 21 distinct transparency requirements derived from GDPR Articles 13 and 14, ensuring comprehensive coverage of regulatory obligations.

(2) We propose a configurable, multi-stage LLM-based annotation pipeline that integrates upstream passage-level classifiers and RAG, enabling scalable and precise word-level annotation of addressed GDPR transparency requirements.

(3) We introduce a two-tiered evaluation methodology that captures both passage-level classification and word-level annotation quality, and apply this framework in a comparative analysis of seven prominent LLMs, providing insights into their performance across annotation tasks.

(4) To support robust empirical evaluation, we implement an automated pipeline for crawling and preprocessing Android app privacy policies, resulting in a large corpus of 703,791 English-language policies and an evaluation dataset of 200 policies with manually curated word-level annotations adhering to the annotation scheme described above.

(5) We evaluate the proposed annotation pipeline on this evaluation dataset as well as the widely-used OPP-115 dataset, highlighting the applicability and generalisability of our approach across different annotation standards.

Table 1: Transparency Requirements derived from GDPR Articles 13 and 14.

	Transparency Requirement	GDPR References
1.	Controller Name	13(1)(a), 14(1)(a)
2.	Controller Contact	13(1)(a), 14(1)(a)
3.	DPO Contact	13(1)(b), 14(1)(b)
4.	Data Categories	14(1)(d)
5.	Processing Purpose	13(1)(c), 14(1)(c)
6.	Legal Basis for Processing	13(1)(c), 14(1)(c)
7.	Legitimate Interests for Processing	13(1)(d)
8.	Source of Data	14(2)(f)
9.	Data Retention Period	13(2)(a), 14(2)(a)
10.	Data Recipients	13(1)(e), 14(1)(e)
11.	Third-country Transfers	13(1)(f), 14(1)(f)
12.	Mandatory Data Disclosure	13(2)(e)
13.	Automated Decision-Making	13(2)(f), 14(2)(f)
14.	Right to Access	13(2)(b), 14(2)(c)
15.	Right to Rectification	13(2)(b), 14(2)(c)
16.	Right to Erasure	13(2)(b), 14(2)(c)
17.	Right to Restrict	13(2)(b), 14(2)(c)
18.	Right to Object	13(2)(b), 14(2)(c)
19.	Right to Portability	13(2)(b), 14(2)(c)
20.	Right to Withdraw Consent	13(2)(c), 14(2)(d)
21.	Right to Lodge Complaint	13(2)(d), 14(2)(e)

2 The GDPR and Transparency

The GDPR establishes a legal framework for data protection and privacy in the European Union, setting stringent requirements for entities processing personal data. A core principle of the GDPR is *transparency*, which mandates that individuals be fully informed about how their data is collected, processed, and shared. To enforce this principle, Articles 13 and 14 GDPR define a structured set of information that data controllers must provide to data subjects. Table 1 summarises these transparency requirements into a comprehensive annotation scheme, categorising them according to their legal basis within the regulation.

A data controller, as defined in Art. 4(7) GDPR, is the entity that determines the purposes and means of processing personal data and is thus responsible for ensuring compliance. This is particularly relevant for digital applications that process user data, where the applications' developers, who are typically also the controllers, must disclose key details to users under Art. 13 GDPR. Accordingly, transparency requirements (Reqs.) 1 and 2 of the proposed annotation scheme address the provision of the controller's name and contact details, as well as those of their designated Data Protection Officer (DPO), where applicable (Art. 37 GDPR).

Beyond contact details, controllers must clearly state the purpose of data processing (Req. 5), as required by Art. 13(1)(c) GDPR. This obligation stems from the principle of purpose limitation (Art. 5(1)(b) GDPR), which dictates that data processing must be limited to explicitly stated, legitimate purposes. Furthermore, controllers must disclose the legal basis for processing (Req. 6), selecting from the grounds listed in Art. 6 GDPR. Common justifications in digital applications include user consent, performance of a contract, and legitimate interest [10]. For instance, third-party tracking may only be conducted based on user consent [20]. If consent is used as a legal basis, Art. 13(2)(c) GDPR further requires that users be informed about their right to withdraw consent in a manner as simple as giving it (Req. 20). Similarly, if data processing is based on the performance of a contract (Req. 12) or legitimate interest (Req. 7), additional contextual information must be provided.

One of the most significant transparency-related challenges arises in the disclosure of data recipients (Req. 10) under Art. 13(1)(e) GDPR. This requirement is particularly relevant in digital applications, where data is frequently transmitted via embedded third-party software libraries [8] that can facilitate extensive data sharing, often outside the direct control of app developers [19].

Additionally, if data is transferred outside the European Economic Area, controllers must inform users about such third-country transfers (Req. 11). Art. 13(2)(f) GDPR further requires the disclosure of automated decision-making processes (Req. 13), particularly those with legal or similarly significant effects (Art. 22 GDPR).

Users must also be informed about their rights under the GDPR (Reqs. 14–21), which include the right to access, rectification, erasure, restriction, objection, portability, and complaint submission (Arts. 15–21 GDPR). Merely listing these rights is insufficient: controllers must explain their implications and provide practical steps for exercising them [7].

While Art. 13 GDPR applies when data is collected directly from the data subject, Art. 14 GDPR imposes additional transparency obligations when data is obtained from third-party sources. The primary distinction between these provisions lies in Reqs. 4 and 8: controllers must disclose the categories of data collected and the source of the data, respectively. These requirements mitigate the risk of users being unaware of processing activities that involve their personal data. In the context of digital applications, where data is typically collected directly from users [30], Art. 13 GDPR is generally the more relevant provision. However, this distinction is secondary, since Art. 13(1)(c) GDPR mandates controllers to describe the purposes and legal basis, which requires the description of the data (and the categories) processed [47].

Failure to comply with transparency obligations can result in administrative fines of up to 4% of a company's annual turnover, as stipulated in Art. 83(5) GDPR. This further underscores the legal and financial significance of ensuring full compliance and the need for robust auditing mechanisms, not only to assist users and regulators, but also the data controllers themselves.

Table 2: Summary of Related Work for Automated Privacy Policy Information Disclosure Annotation.

Year	Paper	Models	Annotation-level	Annotation Basis	Dataset	#Privacy Policies	Evaluation Metrics			
							A	P	R	F1
2014	Zimmeck and Bellovin [48]	NB, Rule classifier	Segment	Privacy Practices†	Web Crawl★ (Alexa)	150	✓	✓	✓	✓
2016	Wilson et al. [42]	SVM, LR, HMM	Manual: Phrase, Automated: Segment	OPP schema†	OPP-115★	115	✗	✓	✓	✓
2018	Liu et al. [25]	CNN, SVM, LR	Segment, Sentence	OPP schema†	OPP-115	115	✗	✓	✓	✓
	Wilson et al. [43]	SVM, CNN, LR	Segment, Sentence	OPP schema†	OPP-115	115	✗	✓	✓	✓
	Harkous et al. [17]	CNN (Hierarchy of classifiers)	Segment	OPP schema†	Web Crawl★, OPP-115	130K, 115	✓	✗	✗	✗
	Tesfay et al. [39]	NB, SVM, DT, RF	Sentence	GDPR Privacy Aspects*	Web Crawl★ (Alexa)	10	✗	✓	✓	✓
2019	Andow et al. [5]	NER and heuristics	Sentence	Contradiction types*	Google Play Store★	11,430	✗	✓	✓	✗
	Story et al. [36]	SVC	Sentence	Custom mobile-specific schema†	APP-350★	350	✗	✓	✓	✓
2021	Bui et al. [9]	BLSTM-CRF, BERT	Phrase	Refined OPP schema†	OPP-115	30	✗	✓	✓	✓
	Thotawaththa et al. [40]	BERT, SVM, NB, BiLSTM	Section	User & Expert-based categories†	Web Crawl★ (App Store)	1,430 apps	✗	✓	✓	✓
	Alabduljabbar et al. [4]	LR, SVM, RF, CNN, DNN, BERT	Segment	OPP schema†	Alexa Top-10K, OPP-115	5598, incl. 115	✗	✓	✓	✓
	El Hamdani et al. [16]	XLNet, CNN, T5-11B	Segment	OPP schema†, GDPR Art. 13, 14*	OPP-115 subset, [24] subset	115, 15	✗	✓	✓	✓
2022	Arora et al. [6]	BERT, M-BERT	Phrase	MAPP schema*	MAPP★	300 annotated, 205,973 in total	✗	✗	✗	✗
2023	Xiang et al. [45]	SRL, NER	Sentence	GDPR Art. 13, 14*	UK Google Play Store★	300 annotated, (205,973 in total)	✓	✓	✓	✓
	Tang et al. [37]◇	GPT-4, ChatGPT, Claude2, PaLM, Llama 2	Sentence, Segment	GDPR Art. 13*, OPP-115 schema†	PPGDPR, OPP-115	304, 115	✓	✓	✓	✓
2024	Huang et al. [18]	GPT-4 Turbo	Phrase	OPP-115 schema derivation†	Web Crawl★ (Vanguard, Russell 3000 ETF)	2,545	✗	✓	✗	✗
	Rodriguez et al. [34]	GPT-4 Turbo, Llama 2	Segment, Paragraph	MAPP-schema*	OPP-115, MAPP	115, 65	✓	✓	✓	✓
2025	**Our Approach**	DeepSeek-R1, Gemma-2, GPT-4o Mini, GPT-4o, Llama-3.3, Mixtral, Phi-4, Qwen-2.5	Word/Phrase	GDPR Art. 13, 14*	Web Crawl★ (Play Store)	200 annotated, 703,791 in total	○	✓	✓	✓

† Not GDPR-related (incl. studies where GDPR relation was not identifiable), ∗ GDPR-related, ◇ Preprint, ★ Created the dataset
A: Accuracy, P: Precision, R: Recall, F1: F1-score, ✓ Included, ✗ Not included, ○ Not applicable

3 Related Work

The emergence of privacy policies as the primary medium for the disclosure of privacy-related data practices, coupled with their complexity, has motivated extensive research efforts aimed at automating their analysis. A persistent challenge in this field is the need for high-quality labelled datasets to serve as ground truth for ML. While expert-annotated datasets are of the highest quality, their creation is resource-intensive and costly. Crowdsourcing has been explored as an alternative [43], but labour costs remain high, resulting in a reliance on a small number of expert-labelled benchmarks. Among these, OPP-115 [42], a phrase-level annotated corpus of 115 privacy policies introduced in 2016, has become a widely used reference [4, 25].

Despite its prominence, however, OPP-115 is not GDPR-aligned, as it predates the regulation and uses a distinct annotation scheme. More recent efforts have thus focused on extending OPP-115 [6] and developing new datasets that explicitly reflect GDPR transparency obligations, particularly those laid out in Articles 13 and 14 [16, 37, 45]. These newer corpora aim for closer alignment with regulatory requirements, but coverage and granularity vary.

Even with the availability of datasets featuring fine-grained manual annotations, such as OPP-115, most automated approaches remain focused on segment- or sentence-level annotation, due in part to model constraints. Early work relied on classical ML models such as Support Vector Machines (SVMs) and Logistic Regression (LR) [42, 48], later giving way to deep learning models like Convolutional Neural Networks (CNNs) [17, 25]. The introduction of transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT), enabled more effective word- and phrase-level annotation [9, 16], though such methods remain less common. Notably, Xiang et al. [45] applied a combination of NLP methods such as Semantic Role Labelling (SRL) and Named Entity Recognition (NER) to sentence-level annotation schemes reflecting GDPR requirements, but their approach does not extend to phrase-level labelling.

More recent advances have seen LLMs explored for privacy policy analysis. For example, Tang et al. [37] employed LLMs for segment- and sentence-level classification on the OPP-115 and PPGDPR [26] datasets, focusing on high-level categorisation rather than detailed word- or phrase-level annotation. Rodriguez et al. [34] systematically evaluated hyperparameter and prompt engineering strategies on the MAPP dataset, highlighting LLMs' sensitivity to configuration but again restricting analysis to segment-level classification.

Table 2 provides an overview of key studies, mapping their methodologies, annotation levels, and datasets. It reflects the field's progression from classical machine learning to deep learning, transformer-based models, and more recently, LLMs, alongside an increasing emphasis on GDPR compliance and annotation schemes based on Articles 13 and 14.

Despite these advances, most prior work remains limited to segment- or sentence-level classification, leaving word- and phrase-level annotation underexplored, particularly for the nuanced transparency requirements of GDPR Articles 13 and 14. Even recent studies with regulatory alignment do not offer approaches capable of systematically extracting and labelling individual transparency-related phrases at scale. To date, the application of transformer-based models and LLMs has largely targeted high-level categorisation rather than precise extraction of legal terms and obligations.

Yet, such granularity is essential for downstream applications like compliance auditing, structured data extraction, and automated policy analysis. Segment-level annotation provides only a coarse overview and cannot capture legally significant distinctions within passages. To reliably assess GDPR compliance, it is crucial to accurately identify specific rights, obligations, and data-sharing practices at the word and phrase level.

4 Research Questions

Addressing the research gap laid out above, this work explores the potential of LLMs to generate fine-grained annotations of privacy policies that explicitly align with GDPR transparency obligations. Unlike traditional ML approaches that depend on large, manually curated datasets for training, LLMs offer the promise of generating structured annotations with minimal supervision. However, their capacity to perform detailed legal annotation remains underexplored. This leads to two central research questions:

RQ1 How can LLMs be leveraged to generate fine-grained word-level annotations in privacy policies that align with GDPR transparency requirements?

RQ2 What evaluation methodology is required to assess the reliability, completeness, and regulatory alignment of such fine-grained LLM-generated annotations?

Section 5 addresses RQ1 by introducing a structured methodology that integrates a modular annotation approach that integrates LLM-based annotators with passage-level classifiers, RAG, and self-correction mechanisms.

Addressing RQ2, Section 6 outlines a two-tiered evaluation framework that assesses annotation quality while accounting for linguistic ambiguity and policy variation. This includes a comparative analysis of seven publicly available LLMs against a manually curated ground truth dataset comprising 200 privacy policies from Android applications.

5 Approach

As outlined above, LLMs demonstrate strong capabilities in natural language understanding, making them a compelling choice for automating the annotation of privacy policies. However, their attention mechanisms [41] and finite context windows introduce a trade-off between input length and annotation granularity. While LLMs perform well on broader tasks such as summarisation or classification over longer texts, fine-grained tasks like word- or span-level annotation require a more localised focus. As input length increases, semantic precision may suffer, reducing annotation accuracy.

Therefore, effective annotation requires a balance: too much text risks diluting attention to individual phrases, while overly narrow inputs may omit essential context. This is especially critical

for privacy policies, where meaning often hinges on surrounding content.

To address this challenge, our approach combines two stages. First, we convert raw HTML documents into structured lists of textual passages that preserve context while remaining within manageable input sizes. Second, we apply an LLM-based annotation pipeline to analyse each passage, producing GDPR-aligned annotations with word-level precision.

5.1 Data Preprocessing

Effective large-scale annotation requires more than managing context length, as it also depends on clean, well-structured inputs.

Despite initiatives advocating for standardised, machine-readable formats [12, 32], privacy policies vary significantly in structure and are typically embedded in noisy HTML documents with extraneous elements such as navigation menus, footers, or advertisements. Additional challenges arise due to misconfigured websites, such as broken links redirecting users to generic landing pages rather than the intended privacy policy.

To address these challenges, our preprocessing pipeline transforms raw HTML documents into structured JSON representations optimised for LLM-based annotation. The pipeline, illustrated in Figure 1, consists of several steps ① to ⑤, which we describe in the following subsections.

5.1.1 Isolate Main Content. We begin by isolating the main content within the given HTML document to ensure that extraneous elements do not interfere with subsequent processing ①. We adopt a rule-based filtering strategy to detect and remove elements unlikely to contribute meaningful content (e.g., *nav*, *footer*, *script*).

Additionally, a rule-based heuristic identifies the primary content container most likely to hold the privacy policy by examining semantic indicators, such as tag names (e.g., *main*, *article*) and relevant attribute patterns. Documents lacking identifiable policy content (e.g., error pages) are discarded.

5.1.2 Apply Filters. Given the scale and variability of input compiled from automated web crawls, our pre-processing pipeline incorporates a range of customisable filters designed to improve efficiency and reliability by reducing the volume of irrelevant or low-quality documents ②.

The preprocessing pipeline is explicitly designed to allow researchers to add and configure additional filters tailored to their specific research needs, such as:

(1) *Language filters:* Ensure that only privacy policies in the desired language(s) are processed.

(2) *Length-based filters:* Exclude documents that are suspiciously short (e.g., fewer than a few dozen words) or excessively long (e.g., erroneous HTML merges or scripts).

(3) *Deduplication filters:* Detect and remove or group duplicate documents to prevent redundancy and reduce computational overhead.

(4) *Other domain-specific filters:* Apply additional constraints, such as filtering based on keyword presence or specific disclaimers.

Section 6.1.1 describes the filter configuration we used to compile the population from which we draw our evaluation sample.
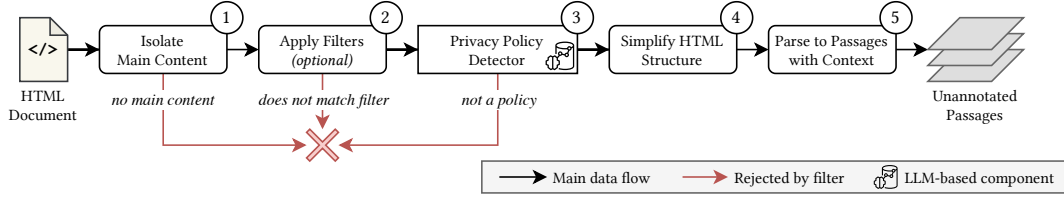
Figure 1: Preprocessing pipeline that parses raw HTML documents into lists of annotation-ready passages.

5.1.3 Privacy Policy Detector. Whereas prior work employed conventional classification approaches such as Logistic Regression [49] and CNNs [24] or smaller language models like RoBERTa [35] to eliminate non-policy documents, Kostina et al. [22] showed that LLMs can outperform conventional classification models in complex text classification tasks.

Based on this observation, we integrate an LLM-based classifier ③ that analyses a short excerpt (~200 words) of a given document and assigns one of three labels: *true* if it is a privacy policy, *false* if it is not, or *unknown* if no determination is possible. Documents labelled *false* or *unknown* are excluded. The prompt used for this task is listed in Appendix B.1.

On a manually labelled sample of 340 randomly selected documents from our evaluation dataset compilation process (see Section 6.1), this approach (using GPT 4.1 as the classifying LLM) achieved an accuracy of 99.7% with no false negatives and a single false positive, where it incorrectly classified general terms of service as a privacy policy.

5.1.4 Simplify HTML Structure. Next, we prepare the document for passage extraction by simplifying the HTML structure to produce a pseudo-HTML representation that retains only meaningful textual elements while preserving hierarchical cues ④.

All superfluous elements and attributes, such as scripts or inline CSS styles, are removed. The Document Object Model (DOM) structure is flattened by unwrapping inline elements (e.g., *spans*) and eliminating redundant nesting of elements. Additionally, nonstandard tags are converted into *divs* for consistency.

5.1.5 Parse Passages. The simplified HTML structure is parsed into a list of passages P, where each element of the simplified DOM is treated as a distinct passage $p \in P$ with text content v_p ⑤.

To ensure that each passage retains essential context, we append a list of contextual elements C_p derived from the simplified HTML structure. These contextual elements $c \in C_p$ include relevant section headings, the preceding passage if the passage belongs to a list or table, and, in the case of table cells, the corresponding column and row headers. This approach enables each passage to be examined and annotated separately while preserving a broader contextual understanding, thus addressing the inherent trade-off between input size and context awareness in LLMs.

For further context, we append the passage's HTML element type e_p. We distinguish four element types in the simplified DOM:

(1) *Headlines:* Typically represented by heading tags (e.g., *h1*, *h2*), which delineate sections or subsections.
(2) *Basic text passages:* Single sentences or paragraphs forming the main body of the policy text, denoted by *p* or *div* tags.

(3) *List elements:* Enumerations or bullet lists that organise disclosures about data processing practices, denoted by *li* tags.
(4) *Table cells:* Used in some policies to summarise data collection or usage in tabular format, denoted by *td* tags.

Consequently, the output of this step is a list of all the privacy policy's passages P where each passage $p \in P$ is defined as a 3-tuple:

$$p = (e_p, C_p, v_p) p \in P \tag{1}$$

This representation preserves local meaning and broader context while keeping inputs LLM-friendly. The resulting passage list P is passed to the annotation pipeline.

5.2 Annotation Pipeline

The annotation pipeline [1] augments the structured passages from the preprocessing stage with GDPR-aligned, fine-grained annotations. As illustrated in Figure 2, the pipeline comprises two layers: the annotation layer generates initial word-level annotations, and the self-correction layer revises and refines them. This section defines the annotation schema and details the function of both layers.

5.2.1 Annotations. Within the context of this research, *annotating* refers to the systematic identification and labelling of specific text spans within privacy policies according to predefined categories, namely legally mandated disclosures under GDPR Articles 13 and 14. An *annotation*, therefore, is a labelled span of text relevant to a GDPR transparency requirement. We represent each annotation as a 3-tuple $a_t = (t, l, b)$, where:

- t is the relevant text span comprising one or more words (e.g., *"your email address"*);
- l is the associated label (e.g., *"Data Categories"*);
- $b \in \{0, 1\}$ indicates whether the statement affirms (1) (e.g., *"we collect X"*) or denies (0) (e.g., *"we do not collect X"*) the addressed requirement.

We define specific annotation guidelines to ensure consistency and precision. General introductions, explanatory content, and references to other sections are excluded, as they do not directly reflect GDPR transparency disclosures. Annotations target the smallest relevant text span that clearly expresses the requirement, including necessary restrictive or defining clauses (e.g., *"your name"*, *"other companies we are affiliated with"*). Irrelevant injected clauses are omitted; thus, annotations may span discontinuous segments (e.g., *"your [...] e-mail address"*). Passages may contain multiple relevant phrases and therefore receive multiple annotations, potentially with the same label. Phrases can also fulfil several GDPR requirements

---

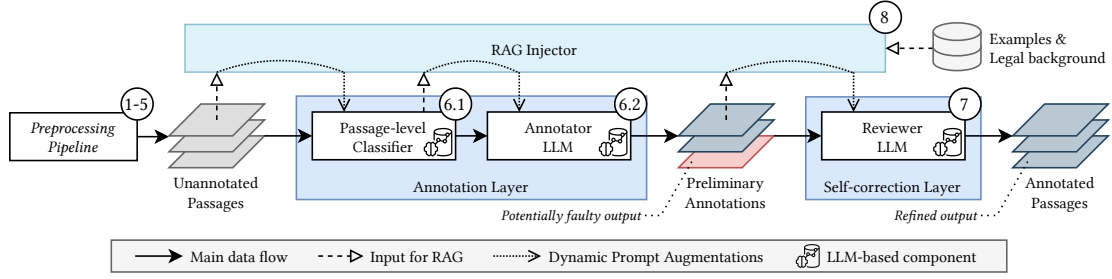[1]https://github.com/tomcory/privacy-policy-annotator

Figure 2: Annotation Pipeline comprising LLM-based annotation- and self-correction layers. The annotation layer combines its core LLM-based annotator with an upstream passage-level classifier to produce annotated passages, which are refined by the self-correction layer's LLM-based reviewer. All LLM-based components are supported by a dedicated RAG injector, which dynamically augments their inputs with suitable examples and legal background.

simultaneously, leading to multiple annotations with different labels. For example, *"promoting our business through marketing"* may qualify as both a *Processing Purpose* and, if explicitly stated, a *Legitimate Interest for Processing*. In cases of conjunction reduction, where a shared phrase applies to multiple items but is stated only once, each annotation must include the full context. For instance, in *"You have the right to access and delete your data"*, two annotations are required: *"You have the right to access [...] your data"* and *"You have the right to [...] delete your data"*.

5.2.2 Annotation Layer. The annotation layer processes the set of passages P produced by the preprocessing pipeline presented in Section 5.1.

Each passage is processed individually by the annotation layer to ensure that the annotator LLM receives input within its optimal context window, minimising unnecessary information while maintaining sufficient context for accurate annotation. The task of the annotation layer is to generate a set of annotations A_p for a given passage by identifying relevant text spans $t \in v_p$ according to the annotation guidelines outlined in Section 5.2.1.

However, text annotation is inherently a compound task for machine learning models, as it requires the models to identify which labels are addressed by the passage and localise the precise words that correspond to each label. By breaking this compound task into distinct passage-level classification and word-level annotation tasks, we reduce the complexity of each sub-task. Previous research has mainly focused on conventional classifiers for the first sub-task, as exemplified by Xiang et al. [45], who developed a sentence-level classifier for GDPR requirements with an overall F1-score of 80%. However, recent work by Kostina et al. [22], as well as our own findings with the privacy policy detector (Section 5.1.3), demonstrate that LLMs can substantially outperform these conventional approaches on complex classification tasks.

Motivated by these findings, the annotation layer integrates an LLM-based passage-level classifier ⑥.①. This LLM-based classifier predicts the set of labels L_p relevant to each passage such that

$$L_p = \{ l \in \mathcal{L} \mid f(p, l) = 1 \} \tag{2}$$

where $\mathcal{L}$ is the set of all labels and $f(p, l)$ is an indicator function that returns 1 if the classifier predicts that l is present in p.

Given L_p, the word-level annotator model ⑥.② focuses solely on identifying the exact spans for each predicted label rather than performing full multi-label classification from scratch. This modularisation narrows the search space for annotation with the goal of reducing LLM hallucination and supporting fine-grained span extraction even in complex legal texts.

Each identified span is assigned an annotation $a_t = (t, l, b), a_t \in A_p$, where l denotes the addressed GDPR transparency requirement and b indicates whether the phrase affirms or negates the requirement. This process transforms each passage into an annotated 4-tuple:

$$p' = (e_p, C_p, v_p, A_p), \quad p' \in P' \tag{3}$$

where A_p represents the set of generated annotations for the passage and P' is the set of preliminarily annotated passages, which is passed to the subsequent self-correction layer.

5.2.3 Self-correction Layer. The self-correction layer refines the preliminary annotations generated by the annotation layer by systematically reviewing the set of annotations A_p for each passage $p' \in P'$. This step is necessary to mitigate potential errors in label assignment, phrase identification, and annotation completeness.

The self-correction layer leverages an LLM-based reviewer ⑦, which may perform any number of the following actions to revise A_p:

(1) Modify an existing annotation's label l or boolean value b if errors or inconsistencies are detected.
(2) Remove an annotation if it is determined to be incorrect or irrelevant based on the context of the passage.
(3) Add a new annotation a_t to a text span t that meets the criteria for annotation outlined in Section 5.2.1.
(4) Correct an inaccurately identified span (i.e. a span where either too few or too many words are included) by applying (2) to remove the inaccurate annotation and (3) to add a new annotation to the correct span.

Following this review process, each passage's set of annotations A_p is updated to a revised set A'_p, transforming the passage representation into:

$$p'' = (e_p, C_p, v_p, A'_p), \quad p'' \in P'' \tag{4}$$

where P'' is the refined set of annotated passages, which serves as the final output of the annotation pipeline (see Appendix A) and forms the basis for downstream evaluation and analysis.

Like all generative ML models, LLMs have a tendency to hallucinate incorrect information [28]. To mitigate the impact of such hallucinations, we implement guardrails within the annotation and self-correction layers that prevent them from generating annotations that pertain to made-up labels or text spans. Invalid labels are replaced with the valid label with the highest SBERT cosine similarity. If the cosine similarity is too low, the annotation is discarded. Annotations with invalid text spans are discarded outright.

5.2.4 Prompt Engineering. When developing an LLM-based annotation system such as the proposed approach, careful attention must be paid to the design and formulation of prompts, as their quality and clarity directly impact the accuracy, consistency, and reliability of the model outputs [34].

With this in mind, we adopted an iterative prompt development approach for our annotation pipeline, starting with simple prompts instructing the LLMs to "annotate or review a given passage based on the transparency requirements mandated by GDPR Articles 13 and 14". The prompts then underwent an iterative refinement process based on evaluations of the LLM-generated output on a sample (n=20) of privacy policy passages and resulting tweaks to address observed errors and inaccuracies. Through this feedback loop of manual evaluation and tweaking, the prompts evolved to encompass several components, each serving a distinct purpose:

(1) *Background and Context*: Briefly outlines the regulatory environment to anchor the LLM and its task within the broader legal context.

(2) *Task Definition*: Explicitly defines the annotation task as identifying and labelling specific phrases within privacy policies according to GDPR transparency requirements. This task definition is expanded for the self-correction layer to clarify that the input comprises passages with pre-existing annotations to be reviewed and corrected.

(3) *Label Classes and Legal References*: Enumerates and describes the 21 GDPR transparency requirements presented in Section 2, including specific references to relevant GDPR paragraphs and concise illustrative examples of relevant phrases for each transparency requirement. If the annotation layer is configured to use a passage-level classifier, the RAG injector modifies the task description accordingly and replaces the full list of transparency requirements with the labels predicted by the classifier.

(4) *General Annotation Guidelines*: Provides guidelines outlining which content should and should not be annotated. Initial observations during the iterative prompt development revealed a tendency for LLMs to annotate too eagerly, leading to increased false positives. Thus, the primary objective of these guidelines is to curb this over-eagerness and reduce the incidence of incorrect annotations.

(5) *Linguistic and Grammatical Instructions*: Offers explicit instructions addressing the language-specific annotation criteria outlined in Section 5.2.1, such as handling conjunction reductions, restrictive clauses, and interruptions by irrelevant clauses.

(6) *Additional legal context and background*: Provides relevant legal context, such as GDPR excerpts and recitals.

(7) *Output Format*: Instructs the model on the necessary output content and provides a JSON schema detailing the exact output format to ensure that the generated annotations are directly usable by subsequent processing steps.

The full prompts used for classification, annotation, and self-correction are provided in Appendix B.

5.2.5 RAG Injector. RAG is a technique that enhances the performance of LLMs by supplementing their input with relevant external information retrieved from a curated knowledge base. Instead of relying solely on the information encoded in the model's parameters, RAG dynamically incorporates retrieved content into the model's prompt, thereby improving factual accuracy, contextual grounding, and task-specific performance [23].

To support the annotation and self-correction layers of our pipeline, we implement a custom RAG injector ⑧ designed to augment the LLMs' prompts with supplementary material relevant to the task at hand. Our RAG database consists of two main resource types:

(1) *Legal background*: GDPR excerpts, recitals, and related commentary for each transparency requirement.

(2) *Illustrative examples*: Our implementation uses manually labelled and annotated policy passages that illustrate human-approved phrasing and label assignment. The database is constructed from a 10% sample of our manually curated evaluation dataset (see Section 6.1, with the remaining 90% reserved exclusively for evaluation to prevent data leakage.

Data is stored as structured records linking each example passage to its annotations, associated labels, and corresponding legal references. Embedded representations of the passages are stored in a vector database to support efficient semantic similarity search, leveraging state-of-the-art vector retrieval methods. Our pipeline employs two complementary task-specific retrieval strategies:

(1) *Similarity-based RAG*: For every input, the RAG injector retrieves example passages from the database based solely on semantic similarity to the input passage, without considering label information. This baseline retrieval forms the foundation for prompt augmentation and is applied at all pipeline stages.

(2) *Label-aware RAG*: Builds upon similarity-based retrieval by incorporating label-specific retrieval. For each predicted (or assigned) label associated with a passage by the upstream pipeline, the injector fetches examples and legal background material that are specifically relevant to that label.

Applied retrieval strategies differ between pipeline steps. For the passage-level classifier, only similarity-based RAG is applied, as label predictions are not yet available. For the annotator and reviewer, however, both similarity-based and label-aware RAG are used. The injector retrieves semantically similar manually annotated passages from the vector database and fetches additional annotation examples and legal background for each relevant label. The retrieved materials are compiled into an augmented prompt, which is passed to the LLMs and helps align model outputs with annotation requirements and regulatory expectations.

6 Evaluation

To assess the performance of our annotation approach, we conduct two experiments using datasets with different annotation schemes:

- *Experiment 1* evaluates our approach when applied to the GDPR transparency annotation scheme introduced in Section 2, using a manually curated dataset of 200 annotated privacy policies.
- *Experiment 2* evaluates the generalisability of our approach by applying it to the OPP-115 dataset and annotation scheme.

In both experiments, we compare different pipeline configurations, ranging from a standalone annotator LLM to full pipeline configurations that include the RAG injector and the annotation layer's passage-level classifier. This ablation setup allows us to quantify the contribution of these supporting components to the overall performance.

This section details the dataset construction, manual review procedure, and our two-tiered evaluation methodology which assesses performance at passage-level and span-level granularity. It also outlines our selection of LLMs used for the evaluation and presents comparative performance results across the selected LLMs.

6.1 Evaluation Dataset

A robust ground truth is essential for evaluating our annotation approach when applied to the GDPR transparency requirements introduced in Section 2. However, as discussed, existing datasets use annotation schemes that are incompatible in granularity and label structure. To address this, we compile a dataset of privacy policies with manually curated annotations adhering to our annotation scheme.

Given the evolving nature of privacy policies that is driven by changing regulations and best practices, we adopt a two-stage strategy to construct an up-to-date and representative evaluation dataset. First, we compile a fresh large-scale corpus of privacy policies rather than rely on corpora such as APP-350 [36] or PrivaSeer [35] that are generally a few years old. From this fresh corpus, we draw a representative sample that captures diverse writing styles and structural patterns, enabling a broad and realistic assessment of our approach's performance.

6.1.1 Population. When determining the source population for our evaluation sample, the privacy policies of mobile applications are particularly relevant. This is primarily due to the nature of mobile devices, such as smartphones and smartwatches, which are equipped with numerous sensors and operate within data-rich environments. This is coupled with the fact that mobile applications run as native software on devices, granting them greater access to personal and sensitive information from these sources than websites running in sandboxed browser environments [21]. Consequently, we can reasonably expect their privacy policies to contain more disclosures addressing GDPR transparency requirements [11].

Another major advantage of leveraging mobile applications is the structured availability of privacy policies on app distribution platforms. The iOS App Store and Google Play (also referred to as the Play Store), the predominant digital distribution services within the iOS and Android ecosystems, mandate that developers publish links to their privacy policies on their apps' store pages.
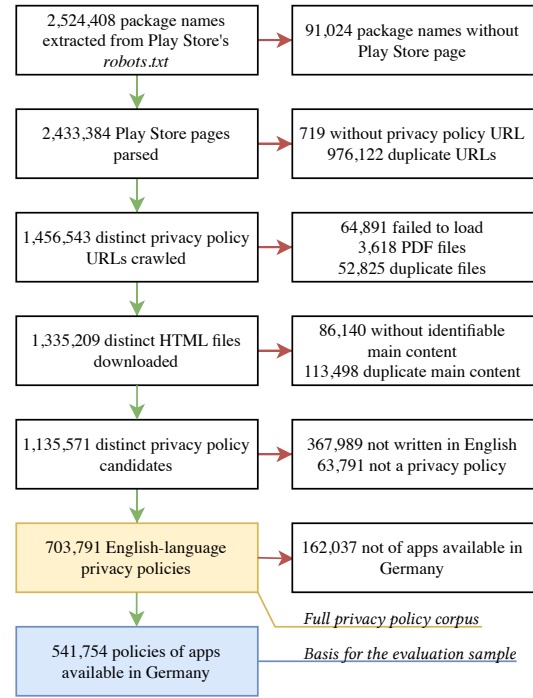

Figure 3: Corpus compilation process, showing the output of each step leading to the corpus of privacy policies used as the population from which we draw our evaluation sample.

Building on this observation, we developed a custom web crawler to automate the retrieval process of Android app privacy policies. The crawler first extracts a comprehensive list of app identifiers from the Play Store's sitemap. It then crawls each app's Play Store page to extract metadata, including the provided privacy policy URLs. Subsequently, it retrieves and downloads the privacy policies, which are passed to our preprocessing pipeline (see Section 5.1).

Figure 3 illustrates the dataset compilation process, initiated in late December 2024. Of the 2,524,408 unique package names extracted from the sitemap, 2,433,384 resolved to active Play Store pages. Parsing these pages resulted in 1,456,543 distinct privacy policy URLs. Following document crawling, deduplication, filtering of non-English documents using LANGID.PY [27], and exclusion of texts not recognised as privacy policies by the LLM-based privacy policy detector (Section 5.1.3), we compiled a corpus of 703,791 distinct English-language privacy policies.

Since our primary objective is to evaluate the annotation of GDPR transparency requirements, we further refine our population to include only privacy policies of applications available for download in Germany, the most populous country in the European Union and the country where the crawl was conducted, yielding 541,754 relevant documents. This final population serves as the basis for our evaluation dataset.

6.1.2 Sample. Having identified a suitable population of privacy policies, we proceed with compiling our evaluation sample.
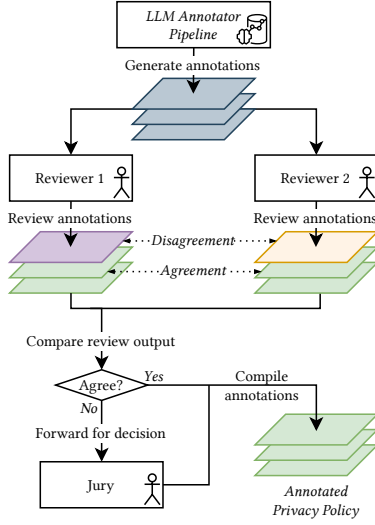
Figure 4: Manual review process.

Privacy policies exhibit substantial variability in length, structure, and language complexity. While some policies concisely address only essential legal requirements, others provide extensive details on data processing practices, legal bases, and user rights. A diverse sample allows us to assess whether our annotation approach effectively handles both concise and verbose documents. However, maximising diversity should not come at the cost of representativeness, as our goal is to infer real-world annotation performance.

To ensure a balanced representation, we apply a diversity-aware sampling strategy that employs k-means clustering to group privacy policies based on their SBERT embeddings. This clustering approach captures structural and linguistic similarities between policies, ensuring that our sample reflects key variations within the dataset. The choice of $k = 4$, selected using the Elbow Method, balances granularity and interpretability, segmenting the dataset without overfitting to minor variations.

The resulting clusters exhibit varying sizes, densities, and internal variability. To mitigate over-sampling from small and dense clusters containing highly similar privacy policies, we apply a weighting formula to determine the relative contribution of each cluster to the final sample. The weight w_c for each cluster $c \in C$ is computed as follows:

$$w_c = H_c \times S_c^2 \times \frac{n_c}{N} \tag{5}$$

$$w_c \in [0, 1], \sum w_c = 1$$

where H_c denotes the Shannon entropy of cluster c, S_c^2 represents its variance, n_c is its size, and N is the total population size. This formulation ensures that clusters with higher diversity and larger sizes contribute proportionally more to the sample.

Given a total sample size of s, the number of privacy policies selected from each cluster (s_c) is determined as:

$$s_c = w_c \times s \tag{6}$$

To further enhance intra-cluster diversity, we apply binned stratified sampling, ensuring a balanced representation of diverse document characteristics from each cluster.

We use a sample size of 200 privacy policies to balance practical constraints with statistical robustness, ensure a sufficiently diverse representation of privacy policies, and remain feasible for manual annotation and evaluation. Appendix C summarises the cluster characteristics of our dataset, as well as the computed sample weights and cluster sample sizes.

6.1.3 Manual Review Process. To establish a high-quality ground truth for evaluating the LLM-generated annotations, we incorporate a manual review process in which human reviewers systematically revise and validate the annotations. This process mirrors the self-correction layer of the annotation pipeline outlined in Section 5.2, but is conducted by domain experts to ensure accuracy and reliability. These reviewers receive detailed briefing sessions and are provided with the same annotation guidelines that were used to instruct the annotation pipeline's LLMs.

To facilitate an efficient and accurate review process, we developed a custom annotation review tool that provides an intuitive graphical interface for human reviewers. The tool highlights annotated phrases in colour-coded formats and allows reviewers to make corrections using simple click-based actions (see Appendix D). This design minimises friction in the review process, allowing reviewers to focus on annotation quality rather than interface complexity.

The manual review workflow is illustrated in Figure 4. Each LLM-annotated passage $p \in P''$ is passed to two human reviewers, who evaluate it independently to produce two revised annotation sets $A_p''^{(1)}$ and $A_p''^{(2)}$. The outcomes of these independent reviews are compared and processed on a per-passage basis as follows:

If the revised annotation sets are identical, i.e., $A_p''^{(1)} \equiv A_p''^{(2)}$, the passage is accepted as final.

If the revised annotation sets differ, i.e., $A_p''^{(1)} \not\equiv A_p''^{(2)}$, the passage is flagged for further review by a jury of experts.

For disputed passages, the expert jury examines the alternative annotation sets and either: (1) selects one of the two proposed annotation sets as the final version, or (2) further revises the annotations to correct inconsistencies or inaccuracies.

This process results in the transformation of LLM-annotated passages into accepted and jury-revised annotated passages p_{gt}:

$$p_{gt} = (e_p, C_p, v_p, A_p'') \tag{7}$$

which are compiled into a final annotated privacy policy document P_{gt} that forms the ground truth for the evaluation methodology presented in the following section.

Using this manual review approach, we employed a team of six domain experts from two universities in Germany and the Netherlands to review the LLM-generated annotations for the sample of 200 privacy policies. The policies were equally distributed among the six experts, with two experts randomly assigned to each policy. After being thoroughly briefed on the task and the applied annotation scheme, the experts reviewed their assigned policies independently of each other.

This initial round of independent reviews resulted in a moderate inter-reviewer agreement (Krippendorff's Alpha) of 0.669,

highlighting the complexity of this annotation task and the importance of robust mechanisms to resolve disagreements. Following the approach laid out above, three of the experts formed a jury to jointly examine and discuss all passages with disagreements, either selecting one of the disagreeing options or, in some cases, opting to overrule both initial reviewers.

Through this process, we compiled a set of 25,260 manually curated annotations across the sample of 200 privacy policies to create the *GDPR-Transparency-200* corpus, which serves as the ground truth for our evaluation.

6.2 Evaluation Methodology

As described at the beginning of this section, we conduct our evaluation using two distinct datasets: the newly created GDPR-Transparency-200 and the OPP-115 dataset. Using the GDPR-Transparency-200 dataset, we assess performance under our primary annotation scheme. In contrast, OPP-115 applies a non-GDPR-aligned annotation scheme with different annotation guidelines, thus enabling us to evaluate the generalisability of our pipeline to varied annotation standards and policy corpora.

As our annotation pipeline comprises multiple components, all of which influence the pipeline's output, we perform an ablation study to evaluate the impact of each component individually and in concert with the surrounding components. Specifically, we measure the performance of:

(1) the standalone annotator with no additional components,
(2) the annotator with downstream self-correction,
(3) the annotator combined with an upstream passage-level classifier and RAG functionality, and
(4) the full pipeline, comprising a passage-level classifier, annotator, and self-correction, all with RAG functionality enabled.

This structured approach allows us to isolate the contributions and interplay of pipeline components systematically and draw conclusions as to which components are especially impactful and which components may benefit from future enhancements.

Given the complexity of our word-level annotation task, traditional multi-label classification metrics are insufficient as they only consider binary presence or absence of labels. Instead, we adopt a two-tiered evaluation framework that measures (1) passage-level performance, examining the passage-level prediction of labels irrespective of annotated text spans, and (2) span-level performance, quantifying the accuracy of annotated text spans associated with each label.

6.2.1 Level 1: Passage-Level Performance.
Passage-level evaluation measures the pipeline's ability to correctly identify applicable annotation labels within passages. This is formulated as a multi-label classification problem, where we extract predicted label sets L_p from annotations A_p and compare them to ground truth labels L_p^{gt} derived from manual annotations A_p''. Both label sets are represented as binary vectors over the complete set of labels, facilitating direct computation of multi-label classification metrics. We calculate *precision*, *recall*, and *F1-score*, and micro-average these metrics across all passages to yield comprehensive performance indicators.

6.2.2 Level 2: Span-Level Performance.
Span-level evaluation assesses how accurately the model identifies the correct text spans

for each label. Given that a label is correctly predicted, we evaluate whether the corresponding span(s) align with those in the ground truth. To quantify this alignment, we compute span similarity using a hybrid metric that combines lexical and semantic similarity:

$$span\theta(t_1, t_2) = \frac{J(t_1, t_2) + cos\theta(t_1, t_2)}{2} \qquad (8)$$

where $J(t_1, t_2)$ denotes the Jaccard similarity between the two spans, capturing structural similarity by measuring their word overlap, and $cos\theta(t_1, t_2)$ represents the cosine similarity of the SBERT embeddings of the text spans, capturing semantic meaning.

Combining word overlap with semantic embedding similarity ensures that spans are evaluated both on structural alignment and meaning, mitigating potential errors due to minor lexical variations. A purely structural approach, such as Jaccard similarity, only measures direct word overlap and does not account for more nuanced differences. For example, if the model omits a qualifier such as *"your"*, this minor deviation still largely preserves the intended meaning. However, if the model mistakenly includes an unrelated word from an adjacent sentence, this represents a more significant error. Both cases might result in similar Jaccard scores despite differing levels of severity. By incorporating semantic similarity through SBERT cosine similarity, we introduce a finer-grained distinction that allows us to penalise severe mismatches more than minor omissions and ensure that the evaluation more accurately reflects the quality of the model's annotations by distinguishing between small acceptable variations and more substantial errors.

Conversely, the inclusion of structural similarity over an exclusively semantic approach provides robustness against overly broad annotations, i.e., annotations of text spans that go beyond the minimal phrase necessary to convey the required information. If we relied solely on semantic similarity, such undesired additions might not meaningfully change the meaning of the annotated span, resulting in a high semantic similarity score despite the significant structural difference.

Using this similarity function, we assess the quality of the LLM-generated set of annotations A_p' by computing pairwise span similarity with the ground truth annotations A_p''. A discrimination threshold τ is applied to map the calculated similarity score to a binary value, determining whether a predicted annotation correctly corresponds to a ground truth annotation.

A *True Positive* occurs when there exists a corresponding ground truth annotation $a_{gt} \in A_p''$ with text span t_{gt} for a given LLM-generated annotation $a_o \in A_p'$ with text span t_o, and vice-versa, such that:

$$TP_o = \{a_o \mid \exists a_{gt} \in: \theta(t_{gt}, t_o) > \tau \wedge l_{gt} = l_o\}, \quad TP_o \subseteq A' \quad (9)$$

$$TP_{gt} = \{a_{gt} \mid \exists a_o \in: \theta(t_{gt}, t_o) > \tau \wedge l_{gt} = l_o\}, \quad TP_{gt} \subseteq A''$$

where $\theta(t_{gt}, t_o)$ denotes the span similarity between t_{gt} and t_o, and l_{gt} and l_o are the labels of the annotations a_{gt} and a_o, respectively.

A *False Positive* occurs when no matching ground truth annotation exists for a given LLM-generated annotation, such that:

$$FP = \{a_o \mid \nexists a_{gt} : \theta(t_{gt}, t_o) > \tau \wedge l_{gt} = l_o\} \qquad (10)$$

A *False Negative* occurs when a ground-truth annotation is not matched by any LLM-generated annotation, such that:

$$FN = \{a_{gt} \mid \nexists a_o : \theta(t_{gt}, t_o) > \tau \wedge l_{gt} = l_o\} \tag{11}$$

Unlike conventional classification tasks, the range of *True Negatives* for phrase-level annotations is vast, as any non-annotated span in a passage could be considered a True Negative. Given that the number of possible spans (i.e. combinations of consecutive and non-consecutive words) in a passage v_p grows combinatorially with the passage length, this results in highly skewed metrics when True Negative-based measures such as accuracy are applied.

We account for this by focusing on metrics that do not rely on True Negatives. Specifically, we employ *Precision*, *Recall*, and F_1-*score* as the primary evaluation metrics to provide a meaningful assessment of the annotation pipeline's performance. Unlike in conventional binary classification tasks, where the baseline for these metrics is typically 0.5 due to the expected value of random guessing, the vast True Negative space in our setting causes the expected value, and thus the effective baseline, to approach zero as the average passage length increases. In extreme cases, a model could either generate no annotations at all or annotate every possible span in a passage for each label, both of which would result in performance scores approaching zero.

Following the definitions of *True Positives*, *False Positives* and *False Negatives* set out above, micro-average *Precision*, *Recall* and F_1-*score* are defined as:

$$Precision = \frac{|TP_o|}{|TP_o| + |FP|} \qquad Recall = \frac{|TP_{gt}|}{|TP_{gt}| + |FN|} \tag{12}$$

$$F_1 = \frac{2}{Precision^{-1} + Recall^{-1}}$$

6.3 Model Selection

To evaluate the effectiveness of our approach across a diverse range of model architectures and capacities, we selected seven prominent LLMs that represent a cross-section of the LLM landscape as of May 2025. Our selection includes proprietary and open-source models of varying size, as well as different architectural paradigms such as decoder-only transformers and multi-modal models. Specifically, the models used are:

- *DeepSeek-R1 (70B)* [15]: A mid-size open-source reasoning model built by distilling the full-size DeepSeek-R1 model's capabilities into LLaMA 3.3 (70B).
- *Gemma 3 (27B)* [38]: The medium-sized open-source flagship model of the Gemma family, optimised for efficiency and performance across diverse tasks.
- *GPT 4.1* [2]: The multi-modal flagship of OpenAI's GPT family of proprietary high-performance models.
- *GPT 4.1 nano* [2]: A more cost-efficient, lightweight variant of GPT 4.1 designed for lower resource demands.
- *LLaMA 3.3 (70B)* [14]: A medium-size open-source model of Meta's LLaMA family, popular for multilingual and reasoning tasks.
- *Phi-4 (14B)* [1]: A compact open-source model, commonly used in settings requiring lower inference cost.
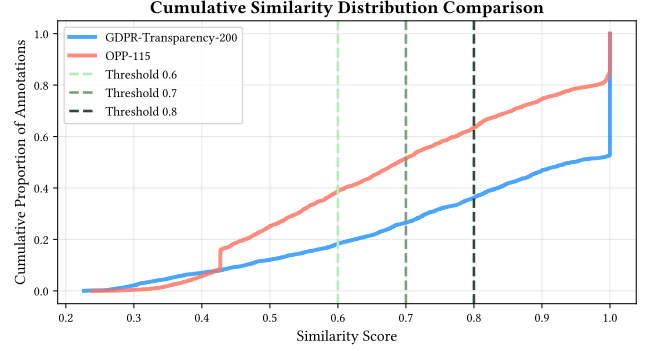

Figure 5: Cumulative annotation span similarity distribution, showing the proportion of annotations with span similarity score $span\theta(t_o, t_{gt})$ below given discrimination thresholds τ across all models and configurations for each dataset.

- *Qwen 3 (32B)* [46]: A medium-size open-source model of the Qwen family, comparable in scale and capabilities to Gemma 3.

By evaluating our approach across this varied selection, we aim to assess both its robustness to different model types and sizes and its scalability across a range of deployment scenarios.

6.4 Evaluation Results

We report annotation results for both GDPR-Transparency-200 and OPP-115, analysing passage-level and span-level performance across pipeline configurations and models. For span-level evaluation, we set the discrimination threshold to $\tau = 0.7$, as this value offers a reasonable compromise that allows minor deviations from the ground truth (especially in cases with complex sentence structure, where identifying the exact set of relevant words for an annotation is challenging even for humans) without being overly permissive and artificially inflating performance scores. Figure 5 visualises the cumulative distribution of annotation span similarity scores across both datasets, with candidate thresholds indicated. As shown, a threshold of 0.7 balances the trade-off between overly lax and overly strict evaluation, providing a consistent basis for subsequent span-level analyses on both datasets.

Table 3 summarises results for all models and configurations. In the baseline setting using standalone annotators, passage-level classification is generally robust: GPT 4.1 achieves the highest precision (0.756) and F_1-score on GDPR-Transparency-200 (0.785), while LLaMA 3.3 delivers the highest recall (0.822). Span-level performance is lower and more varied, with F_1-scores ranging from 0.484 to 0.746. On OPP-115, span-level performance is much lower (highest F_1-score: qwen3:32b with 0.23), as ground-truth annotation span boundaries are often inconsistent and predicted annotation spans rarely match ground truth spans exactly.

Adding the self-correction layer produces mixed results. Stronger models such as GPT 4.1 benefit modestly, but weaker models often see limited gains or even performance drops, suggesting that self-correction can sometimes reinforce existing errors rather than correct them.

Table 3: Evaluation results for the selected models across datasets and pipeline configurations. Best Precision (P), Recall (R), and F_1-scores per evaluation granularity level, configuration and dataset are highlighted in bold. The best performance scores for each dataset across all configurations are highlighted in green .

↓ Model Metric →	GDPR-Transparency-200 Dataset						OPP-115 Dataset					
	Passage-level			Span-level			Passage-level			Span-level		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
(1) Baseline: Standalone annotator (no RAG)												
deepseek-r1:70b	0.719	0.579	0.641	0.568	0.656	0.609	0.740	0.699	0.719	**0.492**	0.121	0.195
gemma3:27b	0.659	0.779	0.714	0.544	0.758	0.634	**0.864**	0.723	0.787	0.408	0.143	0.212
gpt-4.1	**0.756**	0.817	**0.785**	**0.680**	0.824	**0.746**	0.839	0.810	**0.824**	0.460	0.106	0.172
gpt-4.1-nano	0.496	0.659	0.566	0.414	0.583	0.484	0.768	0.709	0.737	0.379	0.098	0.156
llama3.3:70b	0.616	**0.822**	0.704	0.560	0.724	0.632	0.731	**0.844**	0.783	0.400	0.101	0.161
phi4:14b	0.626	0.662	0.643	0.502	0.670	0.574	0.666	0.772	0.715	0.397	0.092	0.150
qwen3-32b	0.657	0.762	0.706	0.515	0.687	0.589	0.776	0.779	0.777	0.404	**0.161**	**0.230**
(2) Standalone annotator with self-correction (no RAG)												
deepseek-r1:70b	0.512	0.424	0.464	0.430	0.512	0.467	0.770	0.685	0.725	**0.504**	0.095	0.160
gemma3:27b	0.764	0.768	0.766	0.606	0.757	0.673	**0.868**	0.730	0.793	0.414	0.133	**0.201**
gpt-4.1	**0.784**	0.799	**0.791**	**0.700**	0.835	**0.761**	0.850	0.803	**0.826**	0.460	0.101	0.166
gpt-4.1-nano	0.460	0.430	0.444	0.390	0.564	0.461	0.771	0.709	0.739	0.381	0.095	0.152
llama3.3:70b	0.570	**0.837**	0.678	0.520	0.713	0.601	0.749	**0.865**	0.803	0.405	0.109	0.172
phi4:14b	0.611	0.559	0.584	0.493	0.611	0.546	0.691	0.782	0.734	0.430	0.098	0.160
qwen3-32b	0.655	0.779	0.712	0.536	0.692	0.604	0.785	0.806	0.795	0.164	**0.153**	0.159
(3) Annotator with upstream passage-level classifier and RAG												
deepseek-r1:70b	0.717	**0.758**	0.737	0.687	0.637	0.661	0.843	0.685	0.756	**0.603**	0.128	0.211
gemma3:27b	0.799	0.719	0.757	0.730	0.731	0.731	**0.899**	0.737	0.810	0.519	0.164	0.249
gpt-4.1	0.856	0.748	**0.798**	0.729	**0.829**	**0.776**	0.876	0.806	**0.840**	0.573	**0.181**	**0.276**
gpt-4.1-nano	**0.857**	0.514	0.643	0.693	0.609	0.648	0.861	0.536	0.661	0.544	0.129	0.208
llama3.3:70b	0.613	0.399	0.483	**0.737**	0.733	0.735	0.787	**0.830**	0.808	0.539	0.160	0.246
phi4:14b	0.749	0.742	0.745	0.557	0.557	0.557	0.829	0.737	0.780	0.569	0.131	0.213
qwen3-32b	0.698	0.770	0.732	0.728	0.736	0.732	0.815	0.775	0.794	0.537	0.161	0.248
(4) Full pipeline: Annotator with upstream passage-level classifier, downstream self-correction, and RAG												
deepseek-r1:70b	0.560	0.819	0.665	0.605	0.615	0.610	0.785	0.758	0.771	0.520	0.113	0.186
gemma3:27b	0.795	0.817	0.805	0.718	0.739	0.728	0.861	0.747	0.800	0.491	0.154	0.234
gpt-4.1	**0.827**	0.834	**0.830**	0.724	**0.844**	**0.780**	0.867	0.834	**0.850**	0.580	0.159	0.250
gpt-4.1-nano	0.700	0.650	0.674	0.689	0.616	0.650	0.810	0.633	0.711	0.544	0.129	0.208
llama3.3:70b	0.583	**0.865**	0.697	0.655	0.740	0.695	0.762	**0.844**	0.801	0.500	0.144	0.224
phi4:14b	0.581	0.836	0.685	0.559	0.605	0.581	0.772	0.796	0.784	0.532	0.126	0.204
qwen3-32b	0.665	0.858	0.749	0.716	0.755	0.735	0.816	0.799	0.808	0.506	0.143	0.223

Adding the upstream passage-level classifier and RAG functionality markedly improves the performance. Decomposing the annotation process and providing targeted context through RAG yields higher passage—and span-level scores for most models, especially on GDPR-Transparency-200. On OPP-115, span-level F_1 scores remain modest due to high variability in ground-truth annotations, but passage-level classification further improves for most models.

The full pipeline, integrating classifier, annotator, self-correction, and RAG, achieves the best overall performance. GPT 4.1 attains the highest passage- and span-level F_1-scores (0.83 and 0.78, respectively) on GDPR-Transparency-200. Improvements from self-correction remain inconsistent for weaker models, however, with three out of seven models showing lower span-level performance that in the configuration without self-correction.

A per-label analysis reveals that models perform best on frequently occurring, short-phrase requirements such as *Controller Name* or *Data Categories*. Performance declines on longer, complex categories (e.g., *Data Retention Period*), and models often confuse closely related labels such as *Processing Purpose* and *Legitimate Interests for Processing*, particularly when policy texts lack clear distinctions. All models are prone to occasional label and text span hallucinations, although this is especially pronounced in smaller models and configurations without RAG.

When comparing datasets, passage-level classification performance is higher for OPP-115, likely due to its simpler annotation label scheme, while span-level evaluation is hindered by inconsistent annotation spans in the ground truth. This effect is clearly visible in Figure 5, where the OPP-115 curve rises steeply, indicating a larger share of low-similarity predictions, whereas GDPR-Transparency-200's curve is more gradual, reflecting closer correspondence between predicted and reference spans. In contrast, GDPR-Transparency-200's stricter annotation guidelines allow higher measured span-level performance despite a lower passage-level performance resulting from the more complex label scheme. This highlights that annotator models are generally capable of identifying the relevant labels for passages, but have difficulty matching the annotated spans of the ground truth if they do not rigorously adhere to strict guidelines addressing linguistic features.

In summary, splitting the annotation task and careful application of RAG and self-correction yield the strongest results, especially for stronger models. However, precise span annotation and accurate attribution of complex label categories remain challenging.

7 Discussion

Our evaluation demonstrates that the proposed pipeline for fine-grained annotation of GDPR transparency requirements in privacy policies provides clear advantages, while also highlighting several limitations and avenues for future research.

A central empirical finding is the benefit of decomposing the annotation task into modular components. Introducing a dedicated passage-level classifier and RAG mechanisms for all LLM-based components substantially improved both passage-level and span-level annotation quality, particularly on the GDPR-Transparency-200 dataset. The classifier effectively focuses the annotator on relevant label sets, reducing both under- and over-annotation, while RAG provides targeted examples and contextual information that improve the localisation of relevant spans. Notably, all evaluated LLMs benefitted from this approach, regardless of their size or architecture, though the magnitude of improvement varied.

At the same time, our results reveal important trade-offs and sources of error propagation. The self-correction layer, while beneficial for stronger models, produced inconsistent gains for smaller or weaker LLMs and occasionally compounded errors from earlier pipeline stages. This underscores the need for careful calibration of reviewer prompts and selective use of self-correction, depending on the underlying model's reliability. Additionally, the pipeline remains sensitive to inaccuracies or ambiguities in classifier predictions and retrieval results, as errors at these stages can propagate and ultimately reduce final annotation quality.

Dataset characteristics also play a critical role in determining pipeline effectiveness. The GDPR-Transparency-200 dataset, with its linguistically precise and fine-grained annotation guidelines, yields higher span-level performance and clearer gains from pipeline modularisation. In contrast, the OPP-115 dataset, with its more heterogeneous annotation spans, proves substantially more challenging for span-level evaluation. This difference, evident both in the cumulative similarity distributions and in performance metrics, suggests that the benefits of automated annotation pipelines are maximised when annotation schemes are rigorous and consistent. Thus, while our approach generalises across differing annotation standards, ground-truth quality and consistency remain critical bottlenecks for reliable automation.

A per-label analysis reveals that LLM-based word-level annotation is most effective for short, well-structured requirements such as *Controller Name* and *Data Categories*, but struggles with labels involving longer, context-dependent spans or ambiguous legal disclosures. Model confusion between related categories (e.g., *Processing Purpose* versus *Legitimate Interests for Processing*) remains a challenge, especially when policy texts lack explicit distinctions. Occasional hallucinations of labels or over-annotation, particularly by smaller models or in the absence of RAG, reinforce the need for robust guardrails and post-processing.

From an application perspective, the pipeline provides a promising foundation for scalable, machine-readable privacy policy annotation. Automated annotation has the potential to support service providers in compliance monitoring, enable regulators to conduct systematic audits, and empower end users through more accurate and structured privacy information. Nevertheless, our results make clear that fully automated solutions are not yet a substitute for expert review, particularly for complex, ambiguous, or rare disclosures. Although it is resource-intensive, the human-in-the-loop review process remains essential for practical deployment in the near future.

Looking ahead, future work should focus on improving the reliability of each pipeline stage, including adaptive retrieval, targeted fine-tuning of models and classifiers, and further optimisation of reviewer workflows. Expanding evaluation to broader datasets, languages, and regulatory frameworks will be critical for establishing generalisability and practical impact. The observed dependence of annotation performance on the quality and consistency of ground-truth datasets also highlights the need for community standards in privacy policy annotation. Furthermore, integrating information sources beyond the privacy policy documents themselves, such as app metadata, presents a promising avenue for improving contextual awareness in automated annotation.

Overall, our findings demonstrate that a modular, classifier- and retrieval-augmented LLM pipeline substantially advances automated privacy policy annotation, especially in settings with rigorous annotation schemes. Realising the full potential of such systems will require ongoing technical refinement and continued integration of human expertise to ensure practical accuracy and reliability.

8 Conclusion

This work introduces a modular, LLM-based pipeline for fine-grained, word-level annotation of privacy policy texts with respect to GDPR transparency requirements. Our empirical findings highlight that decomposing the annotation task by integrating a passage-level classifier, retrieval-augmented generation, and an optional self-correction stage substantially improves annotation quality, especially for structured and well-defined disclosures such as controller information and data categories. The benefits of this modular approach are most evident on datasets with precise and consistent annotation guidelines.

However, the evaluation also reveals persistent challenges for automated systems in handling complex, ambiguous, or context-dependent transparency requirements, particularly where disclosure spans are long or loosely structured. While stronger models benefit most from pipeline modularisation and context augmentation, error propagation and dataset variability remain limiting factors, especially for rare or nuanced categories.

Although the proposed approach reduces manual annotation effort and enhances annotation granularity and consistency, expert human oversight is still necessary for quality assurance, particularly in edge cases and for high-stakes compliance scenarios. Future research should focus on refining each pipeline stage, including adaptive retrieval strategies, targeted model fine-tuning, and expanded integration of human-in-the-loop workflows. Further, the development and adoption of more rigorous and standardised annotation schemes will be critical for enabling scalable and generalisable compliance annotation. Ultimately, such advances will support more effective, large-scale analysis of privacy policies and facilitate the development of practical, trustworthy tools for compliance monitoring and user-facing privacy transparency.

Acknowledgments

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. 2024. Phi-4 Technical Report. arXiv:2412.08905 [cs.CL] https://arxiv.org/abs/2412.08905

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

[3] Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2023. Evolution of composition, readability, and structure of privacy policies over two decades. Proceedings on Privacy Enhancing Technologies 3 (2023), 138–153.

[4] Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ülkü Meteriz-Yildiran, and David Mohaisen. 2021. TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. In Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society. ACM, Virtual Event Republic of Korea, 103–118. https://doi.org/10.1145/3463676.3485608

[5] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In 28th USENIX Security Symposium (USENIX Security 19). USENIX Association, Santa Clara, CA, 585–602.

[6] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar K Bannihatti, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, et al. 2022. A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 5460–5472.

[7] Article 29 Working Party. 2018. Guidelines on Transparency under Regulation 2019/679. Technical Report 17/EN WP260 rev.01. Article 29 Working Party. https://www.edpb.europa.eu/system/files/2023-09/wp260rev01_en.pdf

[8] Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt. 2018. Third Party Tracking in the Mobile Ecosystem. In Proceedings of the 10th ACM Conference on Web Science (Amsterdam, Netherlands) (WebSci '18). Association for Computing Machinery, New York, NY, USA, 23–31.

[9] Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated Extraction and Presentation of Data Practices in Privacy Policies. Proceedings on Privacy Enhancing Technologies 2 (2021), 88–110.

[10] C Castelluccia, S Guerses, M Hansen, JH Hoepman, J van Hoboken, B Vieira, et al. 2017. Privacy and data protection in mobile applications: A study on the app development ecosystem and the technical implementation of GDPR. (2017).

[11] Thomas Cory, Wolf Rieder, and Thu-My Huynh. 2024. A Qualitative Analysis Framework for mHealth Privacy Practices. In 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). 24–31. https://doi.org/10.1109/EuroSPW61312.2024.00010

[12] Lorrie Faith Cranor. 2003. P3P: Making privacy policies more useful. IEEE Security & Privacy 1, 6 (2003), 50–55.

[13] European Parliament and Council of the European Union. 2016. General Data Protection Regulation (GDPR) – Articles 13 and 14. https://eur-lex.europa.eu/eli/reg/2016/679/oj Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

[14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

[16] Rajaa El Hamdani, Majd Mustapha, David Restrepo Amariles, Aurore Troussel, Sébastien Meeùs, and Katsiaryna Krasnashchok. 2021. A combined rule-based and machine learning approach for automated GDPR compliance checking. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (São Paulo, Brazil) (ICAIL '21). Association for Computing Machinery, New York, NY, USA, 40–49.

[17] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In 27th USENIX Security Symposium (USENIX Security 18). USENIX Association, Baltimore, MD, 531–548.

[18] Ziyuan Huang, Jiaming Tang, Manish Karir, Mingyan Liu, and Armin Sarabi. 2024. Analyzing Corporate Privacy Policies using AI Chatbots. In Proceedings of the 2024 ACM on Internet Measurement Conference (Madrid, Spain) (IMC '24). Association for Computing Machinery, New York, NY, USA, 505–515. https://doi.org/10.1145/3646547.3689015

[19] Simon Koch, Manuel Karl, Robin Kirchner, Malte Wessels, Anne Paschke, and Martin Johns. 2025. The Impact of Default Mobile SDK Usage on Privacy and Data Protection. Proceedings on Privacy Enhancing Technologies 1 (2025), 808–823.

[20] Konrad Kollnig, Reuben Binns, Pierre Dewitte, Max Van Kleek, Ge Wang, Daniel Omeiza, Helena Webb, and Nigel Shadbolt. 2021. A fait accompli? an empirical study into the absence of consent to third-party tracking in android apps. In Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security (SOUPS'21). USENIX Association, USA, Article 10, 15 pages.

[21] Konrad Kollnig, Anastasia Shuba, Reuben Binns, Max Van Kleek, and Nigel Shadbolt. 2022. Are iPhones Really Better for Privacy? A Comparative Study of iOS and Android Apps. Proceedings on Privacy Enhancing Technologies 2 (2022), 6–24.

[22] Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large Language Models For Text Classification: Case Study And Comprehensive Review. https://arxiv.org/abs/2501.08457 Issue: arXiv:2501.08457 arXiv:2501.08457 [cs].

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33 (2020), 9459–9474.

[24] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The Privacy Policy Landscape After the GDPR. Proceedings on Privacy Enhancing Technologies 1 (2020), 47–64.

[25] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards automatic classification of privacy policy text. School of Computer Science Carnegie Mellon University (6 2018).

[26] Shuang Liu, Baiyang Zhao, Renjie Guo, Guozhu Meng, Fan Zhang, and Meishan Zhang. 2021. Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13. In Proceedings of the Web Conference 2021. 2154–2164.

[27] Marco Lui and Timothy Baldwin. 2012. langid.py: an off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations (Jeju Island, Korea) (ACL '12). Association for Computational Linguistics, USA, 25–30.

[28] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[29] Victor Morel and Raúl Pardo. 2020. SoK: Three Facets of Privacy Policies. In Proceedings of the 19th Workshop on Privacy in the Electronic Society (Virtual Event, USA) (WPES'20). Association for Computing Machinery, New York, NY, USA, 41–56.

[30] Trung Tin Nguyen, Michael Backes, and Ben Stock. 2022. Freely Given Consent?: Studying Consent Notice of Third-Party Tracking and Its Violations of GDPR in Android Apps. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. ACM, New York, NY, USA, 2369–2383.

[31] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, Serge Egelman, et al. 2019. On the ridiculousness of notice and consent: Contradictions in app privacy policies. In Workshop on Technology and Consumer Protection (ConPro 2019), in conjunction with the 39th IEEE Symposium on Security and Privacy.

[32] Harshvardhan J Pandit, Axel Polleres, Bert Bos, Rob Brennan, Bud Bruegger, Fajar J Ekaputra, Javier D Fernández, Roghaiyeh Gachpaz Hamed, Elmar Kiesling, Mark Lizar, et al. 2019. Creating a vocabulary for data privacy: The first-year report of data privacy vocabularies and controls community group (DPVCG). In On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings. Springer, Cham, 714–730.

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

[34] David Rodriguez, Ian Yang, Jose M. Del Alamo, and Norman Sadeh. 2024. Large language models: a new approach for privacy policy analysis at scale. Computing 106, 12 (Dec. 2024), 3879–3903. https://doi.org/10.1007/s00607-024-01331-9

[35] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6829–6839. https://doi.org/10.18653/v1/2021.acl-long.532

[36] Peter Story, Sebastian Zimmeck, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Natural language processing for mobile app privacy compliance. In *AAAI spring symposium on privacy-enhancing artificial intelligence and language technologies*, Vol. 2. 4.

[37] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. 2023. PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models. https://doi.org/10.48550/arXiv.2309.10238 arXiv:2309.10238 [cs].

[38] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

[39] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. 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* (Tempe, AZ, USA) *(IWSPA '18)*. Association for Computing Machinery, New York, NY, USA, 15–21. https://doi.org/10.1145/3180445.3180447

[40] T.A.I. Thotawaththa, Y.T. Gamage, Damika Gamlath, Wei Chee, and D. Meedeniya. 2021. Automated Categorization of Privacy Policies Based on User Perspective. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, Negambo, Sri Lanka, 54–59. https://doi.org/10.1109/ICIAfS52090.2021.9606158 ISSN: 2151-1810.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[42] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1330–1340. https://doi.org/10.18653/v1/P16-1126

[43] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. 2018. Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations. *ACM Trans. Web* 13, 1, Article 1 (Dec. 2018), 29 pages. https://doi.org/10.1145/3230665

[44] Tong Wu, Qinbo Liu, Binchang Li, Feng Luo, Weilong Li, and Yang Liu. 2023. Detection of Conflicts between APP's Privacy Policy and Actual Behavior: A Security Analysis System. In *2023 7th International Conference on Cryptography, Security and Privacy (CSP)*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 47–53.

[45] Anhao Xiang, Weiping Pei, and Chuan Yue. 2023. PolicyChecker: Analyzing the GDPR Completeness of Mobile Apps' Privacy Policies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (Copenhagen, Denmark) *(CCS '23)*. Association for Computing Machinery, New York, NY, USA, 3373–3387. https://doi.org/10.1145/3576915.3623067

[46] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[47] Gabriela Zanfir-Fortuna. 2020. Article 13 Information to Be Provided Where Personal Data Are Collected from the Data Subject. In *The EU General Data Protection Regulation (GDPR)*, Christopher Kuner, Lee A Bygrave, Christopher Docksey, and Laura Drechsler (Eds.). Oxford University Press: New York, 413–433.

[48] Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *23rd USENIX Security Symposium (USENIX Security 14)*. USENIX Association, San Diego, CA, 1–16.

[49] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proceedings on Privacy Enhancing Technologies* 3 (2019), 66–86.

A   JSON Schema for Annotated Privacy Policies

```
1    {
2      "$schema": "http://json-schema.org/draft-07/schema#",
3      "type": "array",
4      "items": {
5        "type": "object",
6        "properties": {
7          "type": {
8            "type": "string",
9            "enum": ["headline", "list_item", "table_cell", "table_header", "text"],
10           "description": "The type of the passage as defined by the original document's DOM structure."
11         },
12         "context": {
13           "type": "array",
14           "items": {
15             "type": "object",
16             "properties": {
17               "text": {
18                 "type": "string",
19                 "description": "The full text of the context element."
20               },
21               "type": {
22                 "type": "string",
23                 "enum": ["div", "h1", "h2", "h3", "h4", "h5", "h6", "li", "p", "td", "th"],
24                 "description": "The HTML tag type of the context element."
25               }
26             },
27             "required": ["text", "type"]
28           },
29           "description": "A list of relevant context elements (passages) providing contextual information about the passage."
30         },
31         "passage": {
32           "type": "string",
33           "description": "The full text of the passage."
34         },
35         "annotations": {
36           "type": "array",
37           "items": {
38             "type": "object",
39             "properties": {
40               "requirement": {
41                 "type": "string",
42                 "description": "The transparency requirement that is addressed by the annotated phrase."
43               },
44               "value": {
45                 "type": "string",
46                 "description": "The exact phrase that is annotated."
47               },
48               "performed": {
49                 "type": "boolean",
50                 "description": "Indicates whether the phrase addresses the transparency requirement in the positive or negative."
51               }
52             },
53             "required": ["requirement", "value", "performed"],
54             "description": "A list of annotations classifying relevant transparency information within the passage."
55           }
56         }
57       },
58       "required": ["type", "context", "passage", "annotations"],
59       "description": "Represents a passage of a privacy policy as delineated by the original document's DOM structure."
60     }
61   }
```

B   Prompts

B.1   Privacy Policy Detector

```
You are a privacy policy expert. Your task is to analyse the text snippet given below and determine whether it is likely part of a privacy policy or whether it's from another HTML document. General
        terms of service do not count as privacy policies!

Respond with only a single word, omit any additional explanations or context:

'true' if you are sure that the snippet is part of a privacy policy,
'false' if you are sure that it is not, and
'unknown' if there's not enough information to decide.
```

B.2   Passage-level Classifier

```
# Privacy Policy Annotation

You are an expert annotator tasked with classifying privacy policy passages according to their relevance to transparency requirements mandated by the General Data Protection Regulation (GDPR),
        specifically Articles 13 and 14. Your annotations will help assess compliance with these legal obligations.

## Task Definition

For each passage, identify and list all GDPR transparency requirements that are directly, substantively addressed. Output only the names of applicable requirements as a JSON array.

---

## List of GDPR Transparency Requirements

1) **Controller Name**: Disclosure of the name of the data controller (Art. 13(1)(a), 14(1)(a)), e.g. "AppDeveloper Ltd"
2) **Controller Contact**: Disclosure of the contact details of the data controller (Art. 13(1)(a), 14(1)(a)), e.g. "email(at)appdeveloper.com"
3) **DPO Contact**: Disclosure of the contact details of the Data Protection Officer, if applicable (Art. 13(1)(b), 14(1)(b)), e.g. "dpo(at)appdeveloper.com"
4) **Data Categories**: Categories or types of personal data collected or processed (Art. 14(1)(d)), e.g. "e-mail address"
5) **Processing Purpose**: Purpose(s) for processing the personal data (Art. 13(1)(c), 14(1)(c)), e.g. "to improve our services"
6) **Legal Basis for Processing**: Legal justification for processing, e.g., consent, contract, legitimate interest (Art. 13(1)(c), 14(1)(c)), e.g. "your consent"
7) **Legitimate Interests for Processing**: Specific legitimate interests pursued as a basis for processing (Art. 13(1)(d)), e.g. "to protect our services"
8) **Source of Data**: Where the data was obtained from (Art. 14(2)(f)), e.g. "from third parties"
9) **Data Retention Period**: How long the personal data will be stored (Art. 13(2)(a), 14(2)(a)), e.g. "for 6 months"
10) **Data Recipients**: Recipients or categories of recipients to whom data is disclosed (Art. 13(1)(e), 14(1)(e)), e.g. "Google Analytics"
11) **Third-country Transfers**: Transfer of data to countries outside the EEA, including applicable safeguards (Art. 13(1)(f), 14(1)(f)), e.g. "United States"
12) **Mandatory Data Disclosure**: Whether the provision of personal data is mandatory, and consequences of not providing it (Art. 13(2)(e)), e.g. "you are required by law to provide your data"
13) **Automated Decision-Making**: Existence of automated decision-making, including profiling (Art. 13(2)(f), 14(2)(f)), e.g. "profile building"
14) **Right to Access**: Data subject's right to obtain access to their personal data (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to access your data"
15) **Right to Rectification**: Data subject's right to rectify inaccurate personal data (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to correct your data"
16) **Right to Erasure**: Data subject's right to have personal data erased ("right to be forgotten") (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to delete your data"
17) **Right to Restrict**: Data subject's right to restrict processing (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to restrict processing"
18) **Right to Object**: Data subject's right to object to processing (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to object to processing"
19) **Right to Portability**: Data subject's right to receive data in a portable format and transfer it (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to receive your data"
20) **Right to Withdraw Consent**: Data subject's right to withdraw consent at any time (Art. 13(2)(c), 14(2)(d)), e.g. "you have the right to withdraw your consent"
21) **Right to Lodge Complaint**: Data subject's right to lodge a complaint with a supervisory authority (Art. 13(2)(d), 14(2)(e)), e.g. "you have the right to lodge a complaint"

---

## Instructions

- For each passage, determine which (if any) of the 21 GDPR transparency requirements listed above are **directly, substantively** addressed.
- **Do not** make up labels that are not listed below.
- Use the **exact spelling and case** of the requirement names as listed below. Do not modify or abbreviate them.
- Assign a requirement **if and only if** the passage contains clear, specific information about that requirementâĂŤeven if the statement is negative (e.g., "We do not transfer data outside the EEA" âĘŠ
        "Third-country Transfers").
- **Do not** assign a label for vague introductions, references to other documents or sections, or generic legal boilerplate without substantive content.
- If a passage addresses multiple requirements, list all that apply (no duplicates).
- If none are addressed, output an empty array: `[]` (do **not** output the word "None").
- Do **not** extract or output any text spans, explanations, or commentary. Only output the list of requirement names.

---

## Output Format

Format your label predictions as a list. Your output must be JSON following the provided schema. Do not output any additional explanations, text, or commentary.
```

B.3   Annotator LLM

```
# Privacy Policy Annotation

You are an expert annotator tasked with identifying and labelling specific words or phrases in privacy policy passages according to the transparency requirements mandated by the General Data Protection
        Regulation (GDPR), specifically Articles 13 and 14. Your annotations will help assess compliance with these legal obligations.

## Task Definition

An upstream classifier has already determined the relevant GDPR transparency requirements for the given passage. For each transparency requirement, identify and annotate all phrases that address it.

## Transparency Requirements

Here's the full list of transparency requirements. Use this full list only for context - you will be instructed exactly which requirements to annotate further below:

1) **Controller Name**: Disclosure of the name of the data controller (Art. 13(1)(a), 14(1)(a)), e.g. "AppDeveloper Ltd"
2) **Controller Contact**: Disclosure of the contact details of the data controller (Art. 13(1)(a), 14(1)(a)), e.g. "email(at)appdeveloper.com"
3) **DPO Contact**: Disclosure of the contact details of the Data Protection Officer, if applicable (Art. 13(1)(b), 14(1)(b)), e.g. "dpo(at)appdeveloper.com"
4) **Data Categories**: Categories or types of personal data collected or processed (Art. 14(1)(d)), e.g. "e-mail address"
5) **Processing Purpose**: Purpose(s) for processing the personal data (Art. 13(1)(c), 14(1)(c)), e.g. "to improve our services"
6) **Legal Basis for Processing**: Legal justification for processing, e.g., consent, contract, legitimate interest (Art. 13(1)(c), 14(1)(c)), e.g. "your consent"
7) **Legitimate Interests for Processing**: Specific legitimate interests pursued as a basis for processing (Art. 13(1)(d)), e.g. "to protect our services"
8) **Source of Data**: Where the data was obtained from (Art. 14(2)(f)), e.g. "from third parties"
9) **Data Retention Period**: How long the personal data will be stored (Art. 13(2)(a), 14(2)(a)), e.g. "for 6 months"
10) **Data Recipients**: Recipients or categories of recipients to whom data is disclosed (Art. 13(1)(e), 14(1)(e)), e.g. "Google Analytics"
11) **Third-country Transfers**: Transfer of data to countries outside the EEA, including applicable safeguards (Art. 13(1)(f), 14(1)(f)), e.g. "United States"
12) **Mandatory Data Disclosure**: Whether the provision of personal data is mandatory, and consequences of not providing it (Art. 13(2)(e)), e.g. "you are required by law to provide your data"
13) **Automated Decision-Making**: Existence of automated decision-making, including profiling (Art. 13(2)(f), 14(2)(f)), e.g. "profile building"
14) **Right to Access**: Data subjectâ€Žs right to obtain access to their personal data (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to access your data"
15) **Right to Rectification**: Data subjectâ€Žs right to rectify inaccurate personal data (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to correct your data"
16) **Right to Erasure**: Data subjectâ€Žs right to have personal data erased ("right to be forgotten") (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to delete your data"
17) **Right to Restrict**: Data subjectâ€Žs right to restrict processing (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to restrict processing"
18) **Right to Object**: Data subjectâ€Žs right to object to processing (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to object to processing"
19) **Right to Portability**: Data subjectâ€Žs right to receive data in a portable format and transfer it (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to receive your data"
20) **Right to Withdraw Consent**: Data subjectâ€Žs right to withdraw consent at any time (Art. 13(2)(c), 14(2)(d)), e.g. "you have the right to withdraw your consent"
21) **Right to Lodge Complaint**: Data subjectâ€Žs right to lodge a complaint with a supervisory authority (Art. 13(2)(d), 14(2)(e)), e.g. "you have the right to lodge a complaint"

## General Annotation Guidelines

- Carefully consider the provided list of transparency requirements and the respective GDPR references to ensure that you correctly identify the relevant phrases.
- Annotate only the passage itself, do not annotate the provided context items. Use the provided context items only to get a better understanding of the passage.
- Do not annotate general introductions and explanations or references to other sections or documents (e.g. "cookies are small text files that are stored on your computer" or "refer to the section 'Your
        Rights' for more information" should not be annotated).
- Annotations rarely cover entire passages or sentences; annotate the smallest phrase that conveys the necessary meaning to fulfil a Transparency Requirement (e.g. in the sentence "We log device
        identifiers.", only annotate "device identifiers" as **Data Category**).
- Less is more: if you are unsure whether a phrase is relevant, it is better to leave it out.
- Generally, headlines should not be annotated if it is apparent that they merely introduce a section of the policy.
- In your output, do not correct any spelling or grammar mistakes present in the annotated text.
- Never make up information that is not present in the text.
- Never make up labels that are not part of the provided list of predicted transparency requirements.

## Linguistic and Grammatical Instructions

- Include restrictive/defining clauses in the annotation (e.g. "**your** name", "**other** companies **we are affiliated with**", "**our** partners").
- A passage may address multiple different transparency requirements. Thus, a passage may have any number of annotations (e.g. "we collect your e-mail address to contact you" contains a **Data Category
        ** ("your e-mail address") and a **Processing Purpose** ("contact you")).
- Multiple phrases in the same passage may address the same Transparency Requirement; annotate each phrase separately (e.g. "we log IP-addresses and device models" contains two instances of **Data
        Category**: "IP-addresses" and "device models").
- A single word or phrase may have multiple annotations (e.g. "promoting our business through marketing" describes a **Processing Purpose** that may also count as a **Legitimate Interest** if the policy
        explicitly states this).
- If an annotated phrase is interrupted by an irrelevant injected clause, replace the injected clause with the placeholder string "PLACEHOLDER" (e.g. "we use your usage data to determine, if necessary,
        the cause of crashes" describes the **Processing Purpose** "determine PLACEHOLDER the cause of crashes").
- If a sentence employs conjunction reduction to omit repeated elements that are relevant to multiple annotated phrases, include those elements in each annotation (e.g. "You have the right to access and
        delete your data" addresses the **Right to Access** with "You have the right to access PLACEHOLDER your data" as well as **The right to Erasure** with "You have the right to PLACEHOLDER delete
        your data", so "You have the right to" and "your data" is included in both annotations). This also applies to shared restrictive/defining clauses (e.g. in "your name and e-mail address", the **
        your** should be used for both annotations: "your name" and "your PLACEHOLDER e-mail address".

## Transparency Requirements to use

Here are the relevant GDPR transparency requirements as predicted by the upstream classifier. ONLY ANNOTATE FOR THESE LABELS, do not annotate for any of the other transparency requirements or make up
        new labels:

{{RAG_LABELS}}

## Supplementary Material

Below are relevant legal background materials to guide your decisions. Use these to inform your annotation choices and ensure alignment with regulatory expectations.

{{RAG_BACKGROUND}}

I repeat, ONLY ANNOTATE FOR THESE LABELS:

{{RAG_LABELS}}

## Output Format

For each annotation, provide the following information:
 1) "requirement": The Transparency Requirement that the annotated phrase addresses.
 2) "value": The annotated phrase itself.
 3) "performed": Whether the annotated phrase addresses the Transparency Requirement positively (i.e. the phrase explicitly states the information) or negatively (i.e. the phrase explicitly states the
        absence of the information).

Your output must be JSON following the provided schema. Do not output any additional explanations, text, or commentary.
```

B.4   Reviewer LLM

```
# Privacy Policy Annotation

You are an expert reviewer tasked with assessing and correcting annotations of specific words or phrases in privacy policy passages according to the transparency requirements mandated by the General
        Data Protection Regulation (GDPR), specifically Articles 13 and 14. Your revised annotations will help assess compliance with these legal obligations.

## Task Definition

For each passage, assess the given annotations and cross-reference them with smallest relevant phrase(s) that address any of the transparency requirements specified below. Add, remove, or correct
        annotations as required, following the definitions and examples provided.

## Transparency Requirements

Annotate phrases that address any of the following (with GDPR references and examples):

1) **Controller Name**: Disclosure of the name of the data controller (Art. 13(1)(a), 14(1)(a)), e.g. "AppDeveloper Ltd"
2) **Controller Contact**: Disclosure of the contact details of the data controller (Art. 13(1)(a), 14(1)(a)), e.g. "email(at)appdeveloper.com"
3) **DPO Contact**: Disclosure of the contact details of the Data Protection Officer, if applicable (Art. 13(1)(b), 14(1)(b)), e.g. "dpo(at)appdeveloper.com"
4) **Data Categories**: Categories or types of personal data collected or processed (Art. 14(1)(d)), e.g. "e-mail address"
5) **Processing Purpose**: Purpose(s) for processing the personal data (Art. 13(1)(c), 14(1)(c)), e.g. "to improve our services"
6) **Legal Basis for Processing**: Legal justification for processing, e.g., consent, contract, legitimate interest (Art. 13(1)(c), 14(1)(c)), e.g. "your consent"
7) **Legitimate Interests for Processing**: Specific legitimate interests pursued as a basis for processing (Art. 13(1)(d)), e.g. "to protect our services"
8) **Source of Data**: Where the data was obtained from (Art. 14(2)(f)), e.g. "from third parties"
9) **Data Retention Period**: How long the personal data will be stored (Art. 13(2)(a), 14(2)(a)), e.g. "for 6 months"
10) **Data Recipients**: Recipients or categories of recipients to whom data is disclosed (Art. 13(1)(e), 14(1)(e)), e.g. "Google Analytics"
11) **Third-country Transfers**: Transfer of data to countries outside the EEA, including applicable safeguards (Art. 13(1)(f), 14(1)(f)), e.g. "United States"
12) **Mandatory Data Disclosure**: Whether the provision of personal data is mandatory, and consequences of not providing it (Art. 13(2)(e)), e.g. "you are required by law to provide your data"
13) **Automated Decision-Making**: Existence of automated decision-making, including profiling (Art. 13(2)(f), 14(2)(f)), e.g. "profile building"
14) **Right to Access**: Data subjectâÄŽs right to obtain access to their personal data (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to access your data"
15) **Right to Rectification**: Data subjectâÄŽs right to rectify inaccurate personal data (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to correct your data"
16) **Right to Erasure**: Data subjectâÄŽs right to have personal data erased ("right to be forgotten") (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to delete your data"
17) **Right to Restrict**: Data subjectâÄŽs right to restrict processing (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to restrict processing"
18) **Right to Object**: Data subjectâÄŽs right to object to processing (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to object to processing"
19) **Right to Portability**: Data subjectâÄŽs right to receive data in a portable format and transfer it (Art. 13(2)(b), 14(2)(c)), e.g. "you have the right to receive your data"
20) **Right to Withdraw Consent**: Data subjectâÄŽs right to withdraw consent at any time (Art. 13(2)(c), 14(2)(d)), e.g. "you have the right to withdraw your consent"
21) **Right to Lodge Complaint**: Data subjectâÄŽs right to lodge a complaint with a supervisory authority (Art. 13(2)(d), 14(2)(e)), e.g. "you have the right to lodge a complaint"


## Review Guidelines

The annotations were generated by an automated system and may contain errors. Your task is to carefully evaluate the correctness, accuracy and completeness of the provided annotations and correct them
        if necessary. If you deem an annotation correct, leave it unchanged. If you find an annotation to be incorrect, incomplete or superfluous, correct or delete it. If you find that any annotations
        are missing, add them.

The given annotations were generated based on the guidelines listed below. Reference these guidelines carefully when revising the annotations. For each annotation, ask yourself:

1) "Is the annotation label valid, i.e. does it appear in the list of 21 transparency requirements given above, or do I need to correct it?"
3) "Does the annotated span actually address the stated transparency requirement, or do I need to remove it?"
2) "Does the annotation cover the correct span or do I need to adjust the annotated span by adding or removing some words?"
4) "Does the annotation fulfil all of the guidelines listed below?"
5) "Are there any other spans that need to be annotated to fully capture all transparency requirements addressed by this passage?"

### General Annotation Guidelines

- Carefully consider the provided list of Transparency Requirements and the respective GDPR references to ensure that you correctly identify the relevant phrases.
- Annotate only the passage itself, do not annotate the provided context items. Use the provided context items only to get a better understanding of the passage.
- Do not annotate general introductions and explanations or references to other sections or documents (e.g. "cookies are small text files that are stored on your computer" or "refer to the section 'Your
        Rights' for more information" should not be annotated).
- Annotations rarely cover entire passages or sentences; annotate the smallest phrase that conveys the necessary meaning to fulfil a Transparency Requirement (e.g. in the sentence "We log device
        identifiers.", only annotate "device identifiers" as **Data Category**).
- Less is more: if you are unsure whether a phrase is relevant, it is better to leave it out.
- Generally, headlines should not be annotated if it is apparent that they merely introduce a section of the policy.
- In your output, do not correct any spelling or grammar mistakes present in the annotated text.
- Never make up information that is not present in the text.
- Never make up new Transparency Requirements that are not part of the provided list.

### Linguistic and Grammatical Instructions

- Include restrictive/defining clauses in the annotation (e.g. "**your** name", "**other** companies **we are affiliated with**", "**our** partners").
- A passage may address multiple different Transparency Requirements. Thus, a passage may have any number of annotations (e.g. "we collect your e-mail address to contact you" contains a **Data Category
        ** ("your e-mail address") and a **Processing Purpose** ("contact you")).
- Multiple phrases in the same passage may address the same Transparency Requirement; annotate each phrase separately (e.g. "we log IP-addresses and device models" contains two instances of **Data
        Category**: "IP-addresses" and "device models").
- A single word or phrase may have multiple annotations (e.g. "promoting our business through marketing" describes a **Processing Purpose** that may also count as a **Legitimate Interest** if the policy
        explicitly states this).
- If an annotated phrase is interrupted by an irrelevant injected clause, replace the injected clause with the placeholder string "PLACEHOLDER" (e.g. "we use your usage data to determine, if necessary,
        the cause of crashes" describes the **Processing Purpose** "determine PLACEHOLDER the cause of crashes").
- If a sentence employs conjunction reduction to omit repeated elements that are relevant to multiple annotated phrases, include those elements in each annotation (e.g. "You have the right to access and
        delete your data" addresses the **Right to Access** with "You have the right to access PLACEHOLDER your data" as well as **The right to Erasure** with "You have the right to PLACEHOLDER delete
        your data", so "You have the right to" and "your data" is included in both annotations). This also applies to shared restrictive/defining clauses (e.g. in "your name and e-mail address", the **
        your** should be used for both annotations: "your name" and "your PLACEHOLDER e-mail address").


## Supplementary Material

Below are relevant legal background materials for the existing annotations' transparency requirements to guide your decisions. Use these to inform your annotation choices and ensure alignment with
        regulatory expectations.

{{RAG_BACKGROUND}}


## Output Format

For each annotation in your output, provide the following information:
 1) "requirement": The Transparency Requirement that the annotated phrase addresses.
 2) "value": The annotated phrase itself.
 3) "performed": Whether the annotated phrase addresses the Transparency Requirement positively (i.e. the phrase explicitly states the information) or negatively (i.e. the phrase explicitly states the
        absence of the information).

Your output must be JSON following the provided schema. Do not output any additional explanations, text, or commentary.
```
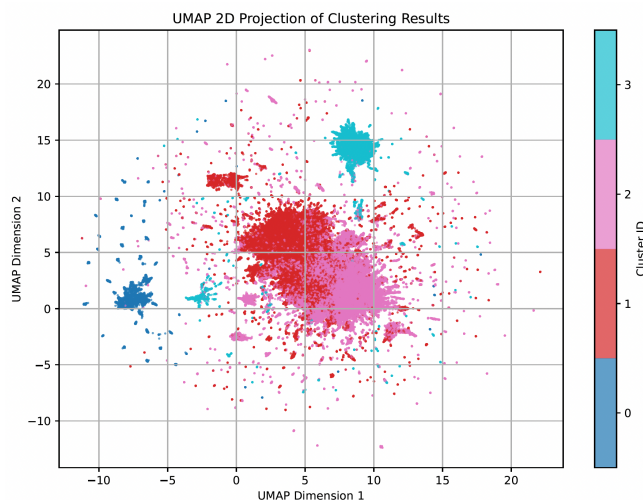
C   Evaluation Sample Compilation

Figure 6: UMAP projection of SBERT embeddings of the 541,754 privacy policies in our corpus, coloured based on their cluster assigned by k-means clustering with k=4.

Table 4: Diversity measures and sample weight of clusters.

c	n_c	H_c	S_c^2	w_c	s_c
0	32,917	1.1613	0.0052	0.0263	6
1	214,103	1.5515	0.0076	0.3314	66
2	237,763	1.5149	0.0087	0.4111	82
3	56,971	1.5810	0.0196	0.2313	46

D   Annotation Review Tool

Figure 7: Annotation review tool used in the manual review process.