Overcoming Language Barriers: Multilingual Analysis of the 2023 Swiss Privacy Law's Impact

Luka Nenadic
ETH Zurich
Zurich, Switzerland
lnenadic@ethz.ch

David Rodriguez
Information Processing and Telecommunications Center
Universidad Politécnica de Madrid
ETSI Telecomunicación, Spain
david.rtorrado@upm.es

Joseph A. Calandrino
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jcalandr@andrew.cmu.edu

Abstract

Policymakers enact and revise privacy laws expecting meaningful benefits for their people in practice. While scholarship has measured the real-world impact of some privacy regulations—the EU and California most notably—limited empirical evidence exists for many of the more than 140 countries that have implemented some form of privacy legislation. Switzerland, a multilingual country bordered almost entirely by EU states, is one such example.

This paper analyzes the extent to which a 2023 alignment of Swiss privacy law with EU privacy regulation affected website privacy policies in Switzerland. To address Switzerland's unique multilingual culture, we develop an LLM-based pipeline that extracts legally relevant information as document-level labels in a single inference without requiring translation. On a benchmark of 120 expert-annotated privacy policies in German, French, Italian, and English, our pipeline achieves F_1 scores above 0.90 for most pairs of languages and legally relevant disclosures.

Applying this pipeline to privacy policies we collected from more than 35,000 Swiss- and EU-facing websites before and after the 2023 privacy law revision, we find significant increases in both mandatory and voluntary disclosures of data subject rights among Swiss privacy policies. In exploring the mechanisms driving increased disclosure rates, we discover heavy use of automated privacy policy generators and find that generated policies are associated with up to 15 percentage points higher disclosure rates. These results provide large-scale empirical evidence of how regulatory change and novel drafting technologies impact the content of privacy policies in a unique multilingual environment.

Keywords

privacy policies, privacy law, LLMs, multilingual analysis

1 Introduction

By 2025, more than 140 countries had implemented data protection regulations [7]. Legislators introduce and amend these laws to meet the privacy expectations and needs of their people. Across countries, privacy laws routinely require transparent disclosure of data processing practices. Evaluating whether privacy laws achieve their intended objectives requires systematic monitoring of their real-world effects. Existing empirical research has focused heavily on the European General Data Protection Regulation (*GDPR*) [21, 24, 28, 59] as well as US state-level privacy legislation—most notably California [35, 75]—while empirical evidence for the privacy laws of other jurisdictions remains strikingly rare. One potential explanation for this scarcity is the substantial measurement challenges associated with systematically assessing legal compliance across jurisdictions, particularly in multilingual settings.

Against this backdrop, this paper evaluates a major revision of Switzerland's Federal Act on Data Protection (*FADP*) that entered into force in September 2023 [71]. The reform aligns Swiss privacy law more closely with the GDPR but maintains differences in the specific requirements for transparency disclosures. Most notably, while Swiss privacy law offers data subjects the same rights as the GDPR—such as the right of data access—it does *not* mandate that these rights be explicitly mentioned in privacy policies (*policies*), contrary to the GDPR. This distinction provides a unique opportunity to examine whether websites' responses to the revision voluntarily extend beyond domestic obligations. If so, this could be interpreted as consistent with the "Brussels Effect" [16], in which European Union (*EU*) regulations like the GDPR exert influence beyond the EU. Overall, the Swiss privacy law revision constitutes a clearly defined regulatory intervention that allows us to examine whether changes in legal requirements are reflected in observable policy disclosures. We thus ask the following research question: *Did the 2023 Swiss privacy law revision significantly impact disclosures in Swiss website privacy policies?*

Switzerland's multilingual setting makes it both an interesting and challenging research environment. To overcome the limitations of traditional methods tailored to the English language (see Section 2.1), we introduce a novel LLM-based pipeline that extracts legally relevant information as document-level labels in a single inference without requiring translation. To evaluate the pipeline's performance, we develop an expert-annotated dataset of 120 policies in German, French, Italian, and English based on an iteratively refined codebook. We annotate seven legally relevant dimensions of transparency disclosures that allow us to monitor the real-world effects of Switzerland's revision as well as potential spillover effects of the GDPR.

Finally, we inspect the role that so-called policy generators (*generators*) played in potentially increasing transparency disclosures. While early forms of generators have been around since the 1970s [18, 27], empirical evidence remains scarce regarding (1) their prevalence (especially in non-English contexts) and (2) the extent to which they can actually improve privacy disclosures. We

Proceedings on Privacy Enhancing Technologies 2026(4), 703–723

https://doi.org/10.56553/popets-2026-0141

hypothesize that generators are primarily used by smaller firms and improve overall disclosure rates by aiding these organizations that may avoid expensive legal professionals. The most-used generators we identify are either free or cost a yearly fee of around $100 [23, 60, 72].

Our *key finding* is that the Swiss revision increased disclosure rates for transparency requirements from both the revised Swiss law and the GDPR, which is consistent with (though not direct evidence of) a Brussels Effect. Based on our novel expert-annotated dataset, we find that our multilingual pipeline leveraging GPT-5 achieves F_1 scores above 0.90 for the vast majority of disclosure dimensions and language pairs (German, French, Italian, and English) that we investigate. We also discover that 18% of the analyzed local Swiss websites *explicitly* use generators, and generator use is correlated with increases in disclosure rates of up to 15 percentage points (*p.p.*)—in contrast with prior literature that has questioned the viability of generators (see Section 2.3).

In summary, our paper makes *three contributions*:

(1) We provide empirical evidence of the real-world effects of Switzerland's 2023 privacy law reform.
(2) We develop, validate, and publicly release a multilingual pipeline for disclosure assessment of policies, along with an expert-annotated benchmark dataset in four languages.[1]
(3) We report on the proliferation and potential benefits of generators in a multilingual environment.

The remainder of this paper is structured as follows: Section 2 summarizes related work. Sections 3 and 4 introduce the methods and dataset, respectively. Section 5 explores the effects of the revision, and Section 6 examines the impact of generators. We discuss our key findings and limitations in Section 7. Section 8 concludes and considers future work.

2 Related Work

2.1 Automated Analyses of Policies

2.1.1 Traditional Methods for Processing Policies. Early approaches to the automated analysis of policies relied primarily on symbolic methods, processing statements through handcrafted lexicons, grammars, and ontologies [3]. Several representative systems illustrate the potential and limitations of this paradigm. PrivOnto [53] constructed a legal ontology to encode data practices such as collection and sharing, enabling structured queries over annotated corpora such as OPP-115 [78]. PolicyLint [5] models disclosures as combinations of actors, actions, and data types to detect internal inconsistencies, and PoliCheck [6] extended this idea by distinguishing between first- and third-party recipients. Although these symbolic approaches offer interpretability and precision, they require substantial manual effort and are brittle under linguistic variation. Hybrid pipelines subsequently combined rule-based extraction with statistical classifiers, and supervised learning methods—particularly support vector machines and logistic regression—became widely adopted for identifying disclosures of data collection, sharing, and user choices [3, 52, 65]. Systems such as MAPS [82] demonstrate that traditional machine learning can scale to larger corpora. Other work explores completeness assessment, alignment of statements, or clustering of policy segments [20, 43, 48].

Over time, neural architectures began to outperform traditional classifiers in policy analysis. Polisis [34], for instance, introduced an end-to-end deep learning pipeline for segmentation and classification, producing structured representations directly from raw policy text. Subsequent research applied neural models to specialized tasks, including question answering and detection of vague or ambiguous language [42, 61]. Despite improved predictive performance, these approaches remained heavily dependent on annotated corpora and predefined taxonomies, which limit their scalability across jurisdictions and languages. In particular, most available datasets and trained models focus predominantly on English-language policies, limiting their applicability in multilingual contexts.

2.1.2 LLM-Based Policy Analysis. Recent studies suggest that large language models (*LLMs*) can substantially facilitate the automated analysis of policies by enabling task specification through natural-language instructions rather than extensive feature engineering and model tuning. Rodriguez et al. [64] evaluated GPT and LLaMA models, reporting that GPT-4 Turbo, under carefully crafted prompts, achieved F_1 scores above 93% on multi-label classification tasks over benchmark corpora. While MAPS [82] and Polisis [34] previously represented the state of the art in scalability and granularity, LLMs can reach comparable or even higher accuracy with substantially less task-specific engineering. In a similar vein, Tang et al. [73] introduced PolicyGPT, a zero-shot framework that reached 87–97% accuracy across multiple policy classification datasets, highlighting the continued importance of annotated corpora for validation. Goknil et al. [29] further demonstrated that performance varies markedly across prompt design strategies, underscoring the sensitivity of prompt-based approaches.

More recent work has extended LLM-based analysis toward legally grounded compliance assessments. Cory et al. [19] performed fine-grained annotation of GDPR transparency requirements using a two-stage pipeline with self-correction, evaluating eight LLMs over 703k app policies. Complementarily, Xie et al. [80] proposed an LLM-based framework to evaluate policies against 34 clauses derived from the GDPR and US state laws, achieving average F_1 scores of 0.94 and enabling large-scale analysis of thousands of websites. While these studies illustrate how LLMs extend prior work toward more fine-grained and legally grounded compliance assessments at scale, they also highlight important limitations: both are restricted to English-language policies and explicitly identify multilingual evaluation as a key direction for future research.

2.2 Datasets for Policy Analysis

Progress in the automated analysis of policies has been closely tied to the availability of annotated datasets. For early symbolic and statistical approaches, corpora such as OPP-115 [78] and APP-350 [82] provided thousands of labeled statements about data collection and sharing practices. MAPP [8] extended this line to multilingual contexts (English and German), while IT-100 [32] targeted GDPR-specific disclosures of international transfers. These datasets served a dual role: training material for supervised classifiers as well as

[1]The codebook, dataset, and method are available in their entirety at: https://doi.org/10.5281/zenodo.20512191.

benchmarks for comparing different methods. Although LLMs typically no longer require task-specific training on these datasets, they still rely on them as a benchmark for validation [33].

Subsequent resources diversified both task framing and scale. PrivacyQA [61] and PolicyQA [1] entail question-answering tasks, linking the policy text with expert-curated evidence or span-level reading comprehension. Moreover, APPCorp [44] annotated document organization at paragraph level. Longitudinal corpora, such as the Princeton–Leuven dataset [4], assembled nearly one million snapshots across a decade to study the evolution of policies post-GDPR. Other important longitudinal corpora include the work of Linden et al. [41] and Wagner [77].

Recent efforts further align datasets with regulatory and jurisdictional contexts. GDPR-NER [22] annotated European policies with entities defined in the Data Privacy Vocabulary [57] ontology, while C3PA [51] encoded over 48,000 policy segments against the disclosure requirements of the California Consumer Privacy Act (*CCPA*). Marotta-Wurgler and Stein [46] annotated multi-class labels capturing legally salient clause interdependencies. Other corpora expand linguistic and geographic coverage, including a Saudi dataset of annotated Arabic policies [47]; the 100-Platforms corpus [58], which samples major services for high-level data practices; and the longitudinal, multilingual scraper by Bernhard et al. [14]. Together, these datasets illustrate a shift from general-purpose classification resources toward regulatory-aligned, multilingual corpora.

2.3 Automated Drafting of Legal Documents

In an effort to comply with privacy laws, parties (especially resource-constrained ones) may turn to automated policy generators (*generators*) to draft tailor-made policies at a fraction of the cost of lawyers. In legal literature, Betts and Jaep [15] provide a historical overview of the use of generators by lawyers and their clients for various types of legal documents. They also explore the possibility of automation via machine learning. Contreras [18] discusses how generative AI could address the shortcomings in scope and range of traditional generators. Barton and Rhode [12] present lawsuits and regulatory restrictions that consumer-oriented legal services such as generators have faced in the US. Many theoretical contributions from legal scholars often discuss generators under the umbrella term of "document assembly systems" [27, 66].

Empirical data related to generators is scarce. To our knowledge, only two studies have illuminated some key aspects of generator use "in the wild." First, Sun and Xue [70] investigate the output quality of policy generators by drafting a total of thirty policies for three synthetic apps and ten generators. Second, Pan et al. [56] examined nearly 50,000 app policies on Google Play and found that generators have likely been used to generate 20.1% of the policies. Similar to the first study, the authors generate synthetic policies to explore compliance issues in the generated documents.

3 Method

The analysis of policies poses a persistent challenge, as they are lengthy, vague, and full of legal jargon, which complicates automated processing [3, 62]. In Switzerland, like in the EU, this difficulty is compounded by the multilingual environment. As a result, any empirical assessment of disclosures must be able to accommodate this linguistic diversity without introducing biases.

Traditional computational approaches (see Section 2.1) are poorly suited to this task. Symbolic methods and lexicon-based classifiers have been developed almost exclusively for English. Supervised learning approaches depend on annotated corpora that exist primarily in English (e.g., OPP-115) and cannot be directly applied to other languages without extensive re-annotation. Even deep learning models such as Polisis [34] or MAPS [82] rely on English-only training data and fixed taxonomies, reinforcing an anglophone bias [49] that limits their adequacy in the Swiss context.

LLMs provide a more flexible alternative: a single model can be applied across different languages without retraining, avoiding the need for language-specific resources or feature engineering. At the same time, LLMs can capture compliance-relevant disclosures—such as the purposes of data processing—in a consistent manner across languages. Importantly, their performance must still be assessed against independent human-coded data to ensure the accuracy, completeness, and legal relevance of model outputs, as LLMs are not immune to decreased performance in specific languages.

3.1 Codebook and Annotations

In line with the "text-as-data" approach in computational social sciences [31], we constructed a multilingual benchmark based on a systematically developed codebook. The codebook translates statutory obligations from the GDPR and the revised Swiss privacy law (*FADP*) into discrete annotation questions. We use these questions to construct the human annotations that serve as the reference benchmark for evaluation as well as to guide the automated assessment, since the codebook is supplied to the LLM in the prompt.

Since translating complex legal requirements into discrete annotations is complex, we first identified obligations that are reflected in policies and can be assessed objectively. This led us to the information duties under Art. 13 GDPR, which we distilled to the following obligations that are reasonably objective to annotate: the controller's identity (including contact details) and the purposes of processing (Art. 13(1) GDPR), as well as the rights of data subjects (access & rectification, erasure, portability, right to lodge a complaint with a supervisory authority, and transparency when performing automated decision-making; Art. 13(2) and 15–22 GDPR).

While distilling complex legal requirements into discrete annotations is necessary to create scalable compliance proxies, we concede that our procedure cannot estimate disclosure quality. Some policies may, for example, add a user-friendly summary of the data subjects' rights to help users understand and invoke their rights.

Our annotated transparency obligations align with prior legal scholarship [28, see online appendix, p. 17]. We extend this work in two ways: we add the controller's identity and contact information and the purposes of data processing as additional transparency disclosures, and we exclude non-binary disclosures such as the data retention period (Art. 13(2)(a) GDPR). This ensures our questions are both (1) validated by previous work and (2) legally grounded.

While this approach yields fewer annotated questions than Cory et al. [19], our questions still serve as robust and scalable proxies for estimating the effect of the 2023 Swiss privacy revision on Swiss website policies. The smaller set also offers a practical advantage:

Code	Codebook question	Legal basis	Disclosure required[1]	Kripp. α[2]
`ispol`	Is the text a privacy policy?	– (screening)	– (screening)	0.672
`upd`	Date of last update	– (metadata)	– (metadata)	–[3]
`contr`	Controller's identity and contact	Art. 13(1)(a) GDPR /Art. 19(2)(a) FADP	GDPR and FADP	0.672
`purp`	Purposes of processing	Art. 13(1)(c) GDPR /Art. 19(2)(b) FADP	GDPR and FADP	0.825
`rect`	Right of access and rectification	Art. 15–16 GDPR / Art. 25 FADP	GDPR only	0.839
`forg`	Right to erasure	Art. 17 GDPR / Art. 30 FADP	GDPR only	0.929
`port`	Right to data portability	Art. 20 GDPR / Art. 28 FADP	GDPR only	0.868
`comp`	Right to lodge a complaint	Art. 77 GDPR / Art. 49(1) FADP	GDPR only	0.933
`hum`	Automated decision-making	Art. 22 GDPR / Art. 21 FADP	GDPR and FADP	0.825

[1] Requirements from both the GDPR and FADP are italicized in all tables throughout the paper.
[2] Values > 0.80 indicate strong reliability, while values between 0.67 and 0.80 are considered acceptable [39].
[3] Not applicable for metadata item.

Table 1: Codebook questions, legal bases, and inter-annotator reliability metrics (Krippendorff's α in second phase).

we can pass all codebook questions to the model in a single inference pass. This mitigates known performance degradation from longer prompts [40, 80] and substantially reduces costs compared to running multiple inference rounds.

The GDPR's data subject rights do not perfectly match the FADP: most notably, the revised FADP does *not* oblige controllers to inform data subjects about their rights: access & rectification, erasure, portability, and lodging a complaint. For the purposes of our annotations, these differences do not matter: any given policy either mentions a specific right (e.g., "You have the right to delete your data.") or not. Since these disclosures are only required under the GDPR, their explicit mention in Swiss-facing websites is consistent with, though not direct causal proof of, the Brussels Effect.

Table 1 lists the full set of questions and coding instructions. We annotate only the controller's practices as stated in the policy and ignore sections unrelated to Switzerland or the EU (e.g., clauses related to California's privacy law). The unit of analysis is the entire policy: each item is coded "1" if the policy explicitly mentions the corresponding element, "0" otherwise; the "last updated" field is coded as the date provided in the policy text or "NA" if absent.

We developed the codebook iteratively in three phases with three annotators: (1) a lawyer with an advanced law degree; (2) a law student nearing completion of an advanced degree; and (3) a privacy engineer (Ph.D.) with five years of work on GDPR-related topics. Edge cases and instructions were discussed among all three throughout the iterative codebook refinement procedure.

The first phase, including 15 English policies, yielded promising reliability metrics (Krippendorff's α = 0.74–1.0). Disagreements led us to refine the controller information (`contr`), data portability (`port`), and automated decision-making (`hum`) questions. A second phase with 20 policies produced α-values between 0.67 and 0.93. Following standard guidelines [39], values above 0.80 indicate strong reliability, while values between 0.67 and 0.80 are generally considered acceptable.

All retained dimensions reached the acceptable threshold; most exceeded 0.80 (see Table 1).[2] Finally, the updated codebook was applied to 120 policies (30 each in German, French, Italian, and English). Each policy was independently annotated by a single coder proficient in the respective language, a common practice once high agreement has been reached [9, 79], which we did in the second round. The 30 policies per language align with prior work [83], where validation sets typically involve only a few dozen policies, as policy annotations are time-intensive.

To ensure consistency across phases, all policies used in the three annotation rounds were drawn from the customized dataset introduced in Section 4. In short, the final dataset consists of the German, French, Italian, and English policies of 35,000 websites collected in August and October 2023. Random sampling was performed from the August 2023 snapshot, independently for each phase. This approach formally allowed overlaps, though in practice only one repetition occurred. These subsets provided a manageable basis for iterative refinement of the codebook and for establishing a multilingual benchmark to validate the method. It also allowed us to annotate the policies closer to a real-world setting, with most transparency disclosures occurring in a balanced way (45–79 positive instances out of the 120 annotated policies), with the exceptions of the purposes of processing (100 positive instances) and automated decision-making (15 positive instances) disclosures.

Beyond serving as a benchmark for validation, our dataset itself constitutes a contribution. Established resources such as OPP-115 or APP-350 remain valuable references and have been widely used [5, 34, 82]. However, their widespread availability raises uncertainty about whether they were incorporated into LLM pretraining, creating a risk of overfitting when used for validation. Our dataset addresses this concern and provides a novel multilingual legal benchmarking resource. We release the codebook and annotations to support reproducibility and enable further research.

3.2 Method Design

Our automated method applies OpenAI models—as Cory et al. reported their comparatively superior performance for detecting GDPR-compliance in English policies [19, p. 520]—to classify policies according to the dimensions defined in our codebook. Throughout this section, we refer to each of the nine dimensions by the

[2]We dropped one codebook question (related to data protection officer) due to inconsistent terminology across policies.

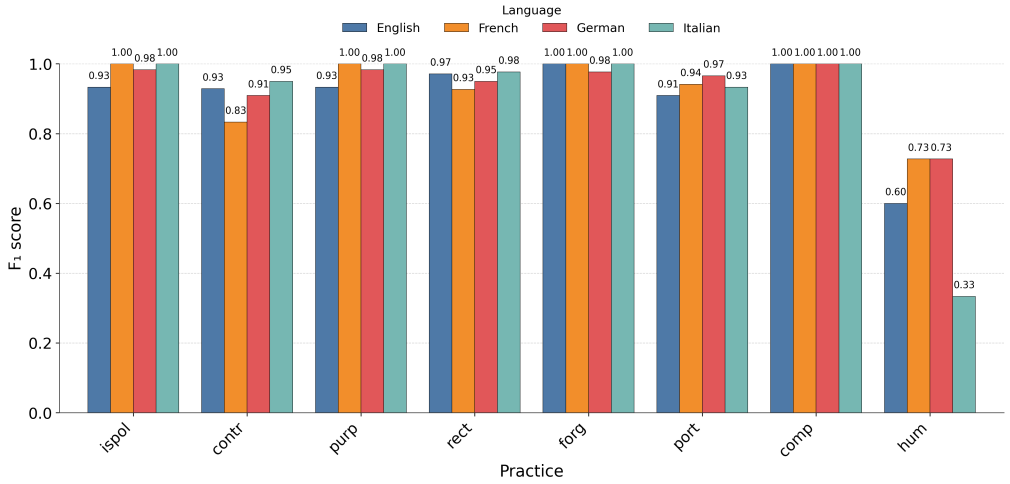


Figure 1: F_1 score of GPT-5 by practice and language (with annotated dataset as benchmark).

abbreviated code as defined in Table 1. For each policy, we issue a single inference request to the model and require it to return a JSON object conforming to a predefined schema. We rely on OpenAI's structured output functionality [54], which enforces the presence and type of all nine fields and prevents the generation of invalid formats or hallucinated attributes. The call is structured with a fixed *system message* specifying the global task and rules, followed by two *user messages*: (1) the codebook, and (2) the raw policy text. We release the prompts and the schema *verbatim* with our artifacts.

The user prompt mirrors the codebook by including each question along with the same decision rules and edge-case clarifications used by annotators. To further reduce ambiguity, we also incorporate few-shot examples demonstrating positive and negative cases for each item. Prior work has shown that such examples significantly improve the reliability of LLM-output in this domain [64], and their inclusion follows standard practice in human codebook development, where annotated examples guide consistent application of coding rules [37, 76, 84]. This alignment also allows direct comparison between manual and automated annotations.

The model processes policies in German, French, Italian, and English without translation. To reduce the risk of the model relying on superficial entity cues (e.g., company name) and to mitigate privacy risks, all identifying information is replaced with placeholders using Presidio [50] prior to inference. Additional global rules are enforced: if `ispol` is false, all other booleans must also be false and the date fields set to "NA." Dates are normalized to the format DD/MM/YYYY for the most recent date mentioned in the policy, or "NA" if none is present. Figure 5 in Appendix A.3 provides an overview of the last updated dates for October policies.

A key feature of our approach is that the model evaluates all disclosure dimensions in a single prompt. Rather than decomposing the task into separate binary classifiers, we treat disclosure assessment as a structured, multi-dimensional extraction problem applied to the policy as a whole. This joint design preserves contextual dependencies between obligations and mirrors the way human annotators apply the codebook to an entire document. In contrast, prior work has typically addressed individual practices in isolation [63, 64, 81]. Beyond efficiency, this unified extraction framework promotes coherent document-level reasoning across legally related dimensions.

We do not apply fine-tuning or task-specific training, and model behavior is determined solely by the fixed prompts and schema. The alias `gpt-5-chat-latest` resolved to `gpt-5-2025-08-07`, which was used for the validation and the large-scale analysis. In repeated validation runs during development, we did not observe instability in the structured outputs for identical inputs, suggesting that schema enforcement and fixed prompting foster stable predictions.

3.3 Method Validation

We validated our method against the independently annotated dataset of 120 policies (30 each in German, French, Italian, and English). This dataset allowed us to directly compare model predictions against human annotations. The final benchmark of 120 policies was treated as a held-out validation set. Prompt design and schema refinement were completed prior to the full validation run, and no prompt adjustments were made based on model performance on this benchmark. This separation reduces the risk of prompt overfitting to the evaluation set.

The annotator with an advanced degree in law reviewed all disagreements between manual annotations and model outputs. This annotator has intermediate Italian proficiency and full professional or native proficiency in the other three languages. When unambiguous annotation mistakes were identified in the human-annotated data (e.g., a policy with the following text: "Art. 22 GDPR: In cases of solely automated processing with legal effects, special rights are granted for data subjects." was erroneously labeled as not mentioning automated decision-making), they were corrected. The full list of corrections is provided in Appendix A.8.[3] For transparency, Appendix A.9 additionally reports robust model performance on the original annotations prior to correction. In our view, the benefits

[3] A total of 21 annotation errors were identified (including one misclassified policy accounting for 7 items), representing less than 2% of all annotations.

of an accurate benchmark for our study as well as for future use outweigh the potential issues of post-annotation corrections.

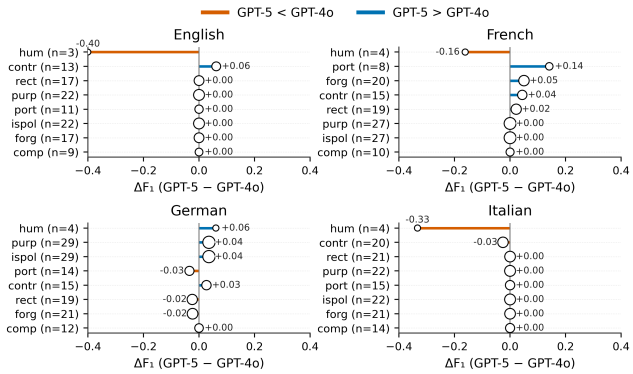
The evaluation metrics were then computed on the corrected benchmark. To assess the stability of these estimates given the sample size per language ($N = 30$), we additionally computed 95% bootstrap confidence intervals for precision, recall, and F_1 by resampling policies with replacement (1,000 resamples, fixed random seed). The resulting intervals are generally narrow across all obligations, indicating stable performance estimates. As expected, uncertainty is substantially larger for hum, which has very few positive cases in the human-annotated dataset (3–4 per language; see Appendix A.2).

We evaluated four OpenAI models—GPT-4o, GPT-4.1, o4-mini, and GPT-5—using identical prompts and the same schema. Figure 1 reports the detailed F_1 results for GPT-5 across languages and practices. Most dimensions reach near-perfect agreement with the human annotations: comp (right to lodge a complaint), ispol (policy detection), purp (purposes of processing), forg (erasure), port (portability), and rect (rectification) all exceed 0.9 in nearly every language. contr (controller information) shows more variability, performing well in English (0.93), German (0.91), and Italian (0.95) but dropping to 0.83 in French. The lowest and most variable scores arise for hum (automated decision-making), with values ranging from 0.73 in German and French to 0.60 in English and 0.33 in Italian. This volatility likely arises from the aforementioned very low number of positive cases in the human annotations, with even a single misclassification substantially affecting the F_1 scores. The low prevalence is likely due to the fact that, in contrast to the other obligations, the requirement to mention hum is *conditional* on the website performing automated decisions. Nonetheless, we must advise caution when interpreting our results for the hum metric.

Results for GPT-4.1 and o4-mini lag behind GPT-5. For GPT-4.1, the largest shortfalls versus GPT-5 occur in port (French, −0.24), contr (German, −0.18), and hum (roughly −0.15 in English, French, and German). Smaller but consistent drops also appear in rect, forg, and purp, while improvements are rare and inconsistent. Results for o4-mini mirror this profile, with similarly large losses in port (French, −0.24) and hum (German, −0.20), together with moderate declines in contr. Gains are again rare, and the model suffers from slower inference speed. Both models show an apparent improvement in hum (Italian, around +0.2), but—as stated—this dimension should be interpreted with caution. Taken together, these patterns make GPT-5 the more reliable and efficient option for evaluating multilingual disclosures of transparency obligations.[4]

We, therefore, narrowed down our comparison to GPT-5 and GPT-4o, as GPT-4o showed the closest overall performance to GPT-5. Figure 2 visualizes their relative performance. The two models are broadly aligned, with most practices showing ΔF_1 scores near zero. GPT-5 offers modest gains—for instance contr in English (+0.06), port in French (+0.14)—while its largest setbacks appear in hum, with −0.40 in English and −0.33 in Italian. Bubble size in the plot reflects the number of positive cases in the human annotations, making clear that the steepest drops occur in the practice with the lowest support. We ultimately selected GPT-5 because it offers a more favorable cost profile and ensures longer-term availability.

[4]We provide the performance metrics of GPT-4.1 and o4-mini in our artifact repository.



Bubble size is proportional to the number of positive human-annotated cases per practice. Labels show ΔF_1 values. Blue bars indicate GPT-5 > GPT-4o, orange GPT-5 < GPT-4o.

Figure 2: Performance difference (ΔF_1) between GPT-5 and GPT-4o across languages and obligations.

4 Dataset

4.1 Initial Dataset

The dataset for this project was created using Bernhard et al.'s [14] scraper. The scraper automatically collects the policies and terms of service of 800,000 websites monthly in 37 languages. We refer to Table 9 in Appendix A.1, drawn and adapted from Bernhard et al. [14, p. 59], for the performance of the scraper at collecting website policies in German, French, Italian, and English. Most notably, the scraper reaches recall scores ranging from 0.78 to 0.91. We also inherited the scraper's functionality of only extracting the single most likely policy per website as a target URL [14, Sections 3.1 and 3.4]. This approach substantially reduced false positives for the vast majority of websites, which is paramount to our analysis. The trade-off, however, is that we do *not* collect multilingual or multijurisdictional versions of a website's policy if such exist.

For the purposes of this paper, we first scraped a sample of 90,000 websites stratified by popularity each in (1) Switzerland and (2) all EU countries, Iceland, Liechtenstein, Norway, and Great Britain. We collected a representative sample of websites with various levels of popularity in August 2023 according to the Chrome User Experience Report (*CrUX* [17]). Due to overlaps between the Swiss group and the non-Swiss group, the procedure resulted in the initial selection of 147,029 websites, of which the scraper could access 144,690 in August and 145,193 in October. For Switzerland, we selected all top 50k websites as well as 20k randomly sampled ones from the 100k-500k and 500k-1M CrUX popularity buckets, respectively. For the non-Swiss group, we collected 500 randomly sampled websites for each popularity bucket (top 1k through 1M-5M, if available) in each country. We control for website popularity rank in our statistical analyses. The initial website budgeting procedure as well as the corresponding CrUX list can be found in our artifact repository.

Each website was scraped once one week before the entry into force of the revision of Swiss privacy law (August 23-28, 2023) and once one month thereafter (October 4-10, 2023). This relatively short time frame increases the likelihood that observed changes occurred due to the revision itself, and not as a result of subsequent

events like court cases. While we acknowledge that longer-term adaptations of websites to the revision may occur more gradually, our conservative approach is consistent with Linden et al.'s [41, p. 50] observation that policy updates clearly clustered within the first month after the GDPR's entry into force.

4.2 Relevant Subset & Website Categories

We utilized only a subset of the initial dataset in our subsequent analyses. To assess whether observed changes are plausibly associated with the entry into force of the revised FADP, we compare disclosure rates across three predefined website groups. Similarly to Peukert et al. [59, p. 752], group assignment is based on (i) top-level domain (*TLD*), (ii) detected policy language using CLD3 [30], and (iii) country-specific traffic popularity buckets (1k, 5k, 10k, 50k, 100k, 500k, 1M, and 5M) from CrUX. We restrict the analysis to policies in German, French, Italian, or English. German, French, and Italian are the official languages of Switzerland,[5] and English is commonly spoken. Websites are categorized as follows:

- `CH`: websites with a `.ch` or `.swiss` TLD that do not appear in any EU-country popularity bucket.
- `CH & EU`: websites with a Swiss TLD that also appear in at least one EU-country popularity bucket.
- `EU`: websites with German, Austrian, French, or Italian TLDs that do not appear in any Swiss popularity bucket.

We refer to the `CH` and `CH & EU` groups jointly as Swiss-facing websites. Websites with generic TLDs (e.g., `.com`) are excluded due to the inability to infer geographic targeting.

Under this design, the `EU` group serves as a quasi-control group, as these websites are not directly subject to the Swiss reform. While indirect spillovers cannot be excluded, any reform-specific disclosure changes should be more pronounced in the Swiss-facing groups than in the `EU` group. We decided to only include websites from Switzerland's neighboring countries in the quasi-control group, because these websites shared a stable privacy environment (GDPR applicable since 2018) and cultural/linguistic ties with Swiss websites, whereas other websites may have introduced noise in our identification method (see Section 5.2) from unrelated events like the California privacy law revision that entered into force in 2023.

Finally, we restrict what we define as valid policies in two ways in order to minimize the occurrence of false positive documents that would undermine our results. We first discard texts that (1) have been classified as a policy with a probability of less than 50% by the classifier introduced by Bernhard et al. [14] or (2) contain less than 200 characters. Our second filtering consists of keeping only the texts that have been classified as a policy according to GPT-5 (see Section 3). In total, GPT-5 discarded 1,013 policy observations (485 in August and 528 in October), which corresponds to 3.6% of the 27,804 policy observations that persisted from the first filtering step. This yields 26,791 valid policy observations across both snapshots.

Figure 3 summarizes the entire construction of the analytical dataset and distinguishes the retained website sample from the valid policy observations used in the subsequent analyses. Each group (`CH`, `CH & EU`, and `EU`) contains more than 5,000 websites, with German as the predominant language across all groups, particularly within `CH` (see Table 11 in Appendix A.4 for the full language distribution).

[5]We do not consider the language Romansh due to its limited availability.

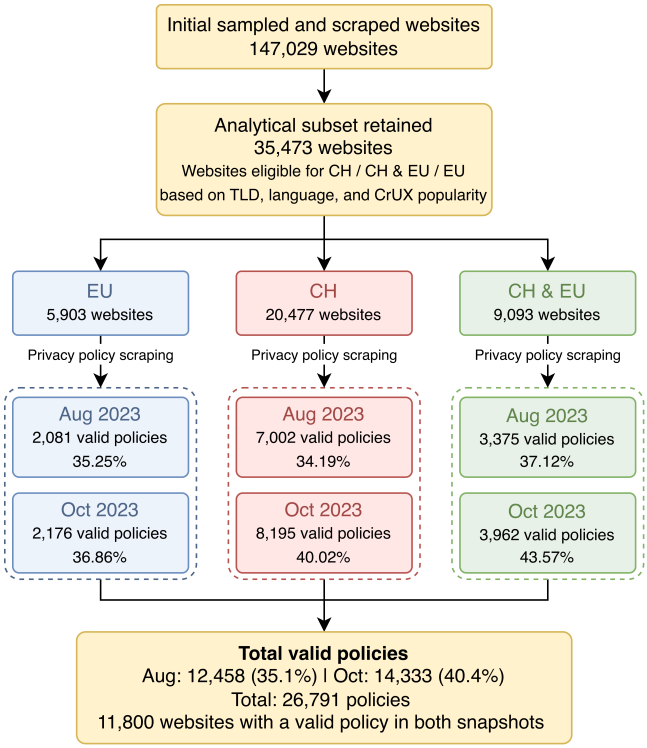


Figure 3: Construction of the analytical dataset.

5 Policies Before and After the Revision

5.1 Descriptive Groups Statistics

Our analysis ultimately builds on a total of 35,473 websites observed at two time points: August 2023 (immediately before the entry into force of the revised FADP) and October 2023 (one month thereafter). The set of websites is held constant across both snapshots; however, the presence of a detectable policy varies between August and October. A valid policy was detected for 12,458 websites in August (35.1%) and for 14,333 websites in October (40.4%). As we discuss below, the Swiss privacy law revision constitutes a likely driver of the increased policy adoption. All subsequent analyses are conducted on the detected policies within each respective snapshot.

The prevalence of detected policies increases between August and October in all groups (see Figure 3). However, this increase is strikingly more pronounced in the Swiss-facing groups (`CH` and `CH & EU`) than in the `EU` group. This suggests differential dynamics around the reform period, indicating that the revised FADP may have driven some Swiss-facing websites without a policy to create one in order to comply with the FADP's disclosure requirements.

We next examine policy length as a coarse proxy for textual changes. Median word counts remain nearly constant in the `EU` group (+14 words), while policies in `CH` and `CH & EU` increase by approximately a median of 300 words (see Table 12 in Appendix A.5). The relative increase is marginally larger in the `CH` group.

Finally, we analyze explicit references to the GDPR and the FADP in policy texts (Table 2). To detect such mentions, we compiled multilingual term lists for both laws (e.g., "DSGVO" for the GDPR, or

	EU	CH	CH & EU
% GDPR August	75.64	62.48	60.71
% GDPR October	75.09	60.77	60.75
% FADP August	7.59	35.83	37.10
% FADP October	7.26	41.12	41.97
Difference GDPR (Δ p.p.)	-0.54	-1.71	+0.04
Difference FADP (Δ p.p.)	-0.33	+5.29	+4.88

Table 2: Mentions of the GDPR and FADP.

"235.1," the FADP's official compilation number; see Appendix A.6 for the full list) and matched them against each policy. Across all groups, references to the GDPR are substantially more frequent than references to the FADP, including within Swiss-facing websites. Between August and October, mentions of the FADP increase in both Swiss-facing groups, whereas GDPR mentions remain stable.

5.2 Disclosure Practices

5.2.1 Snapshot-Level Patterns. We only examine whether a policy explicitly discloses an item as defined in our codebook. We neither assess full legal validity and enforceability of the policies nor the underlying website processing practices. Our analysis is thus confined to observed disclosures in policy texts. For each policy, we first apply the automated classification procedure described in Section 3 to determine whether the policy explicitly discloses each obligation defined in our codebook (Section 3.1). This produces a binary indicator capturing the presence or absence of the respective disclosure for every obligation and policy snapshot. We then compute the proportion of policies disclosing the respective obligation in August and October, and derive percentage-point differences between the two snapshots.

This yields Table 3, which reports disclosure rates in August and October 2023 for each obligation and regulatory exposure group (`EU`, `CH`, and `CH & EU`), together with percentage-point differences between snapshots. Values are computed separately for each snapshot using all valid policies observed at that time.

The descriptive evidence reveals marked differences across regulatory exposure groups. The `EU` group remains largely stable between August and October, with differences for all obligations confined to a narrow range (-0.8 to +0.7 p.p.). In contrast, both `CH` and `CH & EU` groups exhibit consistent positive changes across most obligations. For the `CH` group, increases range from +2.4 to +6.7 p.p. for all obligations except `purp`, with the largest gains observed for `port` and `comp`. A similar but slightly attenuated pattern is present for `CH & EU` websites. Absolute increases tend to be smaller for obligations with higher baseline disclosure levels in August, consistent with ceiling effects in near-saturated categories such as `purp`.

Baseline disclosure in August is also systematically higher for the `CH & EU` group than for `CH` across nearly all obligations, suggesting prior alignment with GDPR disclosure requirements among jointly exposed websites. By contrast, `CH`-only websites start from lower baseline levels and exhibit larger absolute increases by October.

To clarify whether these aggregate patterns are driven by particular website subsets, we further disaggregate October disclosure rates by policy update status, website popularity, and policy language. Policies whose text changed between August and October exhibit the highest average disclosure rate (76.0%), followed by policies that became valid only in October (70.4%), while unchanged policies disclose the least (64.9%).[6] Website popularity is also associated with higher disclosure completeness: websites appearing in a country-specific Top 5k CrUX bucket average 74.9%, compared to 67.5% for less visited websites. Language differences are more limited between German and English policies (68.4% vs. 65.7%), while Italian policies disclose more (70.5%) and French policies less (53.3%); the latter estimates should, however, be interpreted cautiously due to smaller sample sizes. We report the corresponding obligation-level results in Appendix A.7.

Overall, the snapshot-level evidence points to disclosure increases concentrated among Swiss-facing websites. Descriptive comparisons alone, however, cannot establish whether these temporal differences reflect statistically meaningful divergence across the three groups.

5.2.2 Differential Change by Regulatory Exposure. To formally test whether the descriptive differences documented above reflect systematic divergence across regulatory exposure groups, we conduct an inferential analysis that isolates within-website temporal change from compositional effects. Specifically, we restrict the analysis to the balanced panel of websites with an observed policy in both snapshots. The balanced panel comprises $N = 11{,}800$ websites observed in both August and October 2023, including 2,012 `EU`, 6,682 `CH`, and 3,106 `CH & EU` websites.[7] Restricting the analysis to this panel ensures that estimated changes reflect within-website temporal variation rather than shifts in sample composition.

The central question is whether the change in disclosure rates between August and October is significantly larger for the two groups of Swiss-facing websites compared to the EU quasi-control group. To address this question, we estimate, for each disclosure obligation, a difference-in-differences model on the balanced panel using Ordinary Least Squares (OLS). Given the binary disclosure outcome, this specification operates in probability space, allowing the interaction terms to be interpreted directly as differential percentage-point changes. In contrast, non-linear models for limited dependent variables parameterize interaction effects on the odds scale rather than in levels [2]. We confirm that our substantive conclusions do not depend on the regression model choice by estimating a logistic difference-in-differences model in Appendix A.10.

The dependent variable indicates whether the respective obligation is disclosed in a given policy–snapshot observation. The model includes indicators for time (October vs. August), group membership (`EU` as the reference), and interaction terms between time and group. In addition, we include two control variables to account for potential variation in the baselines: website popularity (CrUX popularity rank buckets) and policy language (German, French, Italian, or English). These controls account for systematic differences in disclosure practices associated with market reach and linguistic context rather than the Swiss privacy law revision. Standard errors are clustered at the website level to account for repeated observations of the same policy across snapshots.

[6] We provide all details in Table 16 from Appendix A.7.
[7] Group membership is defined according to baseline (August) regulatory exposure.

	EU			CH			CH & EU		
Obligation	Aug	Oct	Δ p.p.[1]	Aug	Oct	Δ p.p.[1]	Aug	Oct	Δ p.p.[1]
contr	85.8%	86.5%	+0.7	73.0%	76.9%	+3.9	80.5%	83.2%	+2.7
purp	99.3%	99.3%	+0.0	98.7%	98.9%	+0.2	98.4%	98.4%	+0.0
rect	87.7%	87.5%	-0.2	74.6%	77.3%	+2.7	79.9%	82.5%	+2.7
forg	89.7%	89.3%	-0.4	76.8%	79.3%	+2.5	82.5%	84.0%	+1.6
port	72.3%	71.6%	-0.6	46.4%	53.2%	+6.7	54.4%	60.0%	+5.6
comp	73.8%	73.0%	-0.8	44.0%	50.2%	+6.1	52.8%	58.0%	+5.2
hum	26.6%	27.1%	+0.5	16.4%	18.7%	+2.4	23.2%	24.4%	+1.2
Average	76.4%	76.3%	-0.1	61.4%	64.9%	+3.5	67.4%	70.1%	+2.7
Total policies	2081	2176	–	7002	8195	–	3375	3962	–

1. Percentage-point differences (Δ p.p.) are computed within each group using all valid policies observed in August and October 2023, respectively. The composition of policies differs slightly between snapshots.

Table 3: Disclosures by group and obligation (August 2023 versus October 2023).

In this linear specification, the interaction coefficients can be interpreted directly as differential percentage-point changes in disclosure rates relative to the EU baseline. As in any difference-in-differences design, identification relies on the assumption that, absent the Swiss regulatory revision, temporal changes in disclosure would not have differed systematically between Swiss-facing and EU websites. While the two-snapshot structure does not allow direct testing of parallel pre-trends, the empirical stability observed in the EU group between August and October provides a plausible baseline against which differential changes can be evaluated. Moreover, over short time intervals, policies can reasonably be expected to remain stable in the absence of a relevant event, such as a court case or extensive news coverage.

Because we estimate separate models for multiple disclosure obligations, conducting parallel hypothesis tests increases the risk of false positives. We, therefore, control the expected false discovery rate using the Benjamini–Hochberg [13] procedure at $\alpha = 0.05$, applied separately to the two families of interaction terms (CH vs. EU and CH & EU vs. EU).

Table 4 reports the estimated percentage-point changes within each group alongside the difference-in-differences interaction coefficients relative to the EU baseline. The EU group exhibits minimal changes across obligations, supporting the assumption that policies remain largely stable over short time horizons in the absence of a regulatory change. By contrast, the interaction terms for the CH and CH & EU groups are positive and statistically significant for all obligations except purp, indicating that Swiss-facing websites experience significantly larger increases over time than the EU quasi-control group. The strongest differential effects are observed for port and comp, which also display the largest absolute increases in percentage-point terms (see Table 3).

We further examine whether these differential changes are concentrated among particular website subsets by estimating interaction models with an aggregate disclosure index (Appendix A.7, Table 19). The results indicate that website popularity primarily explains disclosure levels: Top 5k websites disclose more in October, but the differential August–October increase is not confined to them. Language mediates the response more substantially. The increase is strongest among German-language policies, which dominate the Swiss-facing sample, while estimates for policies in other languages are smaller and should thus be interpreted with caution. Finally, a descriptive decomposition shows that within-website disclosure changes are concentrated among policies whose text changed between snapshots, consistent with active policy revision during the observation window.

Taken together, the inferential analysis shows that the temporal increases are concentrated among websites directly exposed to the Swiss privacy law revision and significantly exceed the minor changes observed in the EU control group. While alternative explanations, such as other concurrent jurisdiction-specific developments, cannot be entirely excluded, the observed pattern is consistent with a significant increase in both mandatory (according to the revised Swiss privacy law) and voluntary transparency disclosures in the policies of Swiss-facing websites.

6 Privacy Policy Generators

As website operators pondered how to best update their policies to become compliant with Swiss and potentially also EU privacy law, our data offers preliminary evidence that some turned to automated policy generators (*generators*). In the following sections, we describe the key characteristics of generators and the method we used to detect their use to draft policies. We find that generators are not only explicitly used by many Swiss-facing websites (18.2% for October policies in the CH group), but that their use (1) slightly increases over time and (2) is correlated with increases in disclosure rates by up to 15 p.p. for some transparency obligations. We also demonstrate that generators tend to output similar policies with identical clauses, which likely explains the consistently high disclosure rates exhibited by some generators.

6.1 Characteristics

While the various business models and functionalities of generators complicate any simple definition, the core principle of a generator is to *guide users through a questionnaire or similar process that ultimately results in a tailored and (almost) ready-made policy* (see

	Temporal Change (Δ p.p.)			Differential Effect (vs. EU)	
Obligation	EU	CH	CH & EU	CH: DiD Coef. (95% CI)	CH & EU: DiD Coef. (95% CI)
contr	+0.35	+2.38	+1.96	2.03* (1.53–2.53)	1.62* (0.98–2.26)
purp	+0.10	-0.03	-0.13	-0.13 (-0.31–0.06)	-0.23 (-0.46–0.01)
rect	+0.20	+2.05	+2.38	1.85* (1.34–2.36)	2.18* (1.46–2.91)
forg	+0.15	+2.17	+1.71	2.02* (1.51–2.53)	1.56* (0.85–2.27)
port	+0.10	+4.89	+4.93	4.79* (4.15–5.44)	4.83* (3.88–5.77)
comp	+0.00	+4.15	+4.15	4.15* (3.53–4.76)	4.15* (3.25–5.05)
hum	+0.05	+1.69	+1.61	1.64* (0.99–2.29)	1.56* (0.68–2.44)

Percentage-point changes (Δ p.p.) are computed on the balanced panel of websites observed in both snapshots (N = 11,800; 23,600 policy–snapshot observations).
* indicates statistical significance after Benjamini–Hochberg [13] correction at $\alpha = 0.05$, applied separately to the CH vs. EU and CH & EU vs. EU interaction effects.

Table 4: Differential change in disclosure rates between August and October 2023.

generally [15, 18, 27]). We display part of the translated user interface for the most widely used (yet no longer operational) generator in our dataset called "SwissAnwalt" in Appendix A.11. Importantly, users had to agree to cite SwissAnwalt as the source of their policy under threat of litigation before their tailored policy text was generated. We are unclear whether SwissAnwalt actually followed through on litigation threats, but evidence exists that a German law firm offering a generator sends warning letters when its generator is not referenced in the policy [69].

Other generators achieve an even higher level of automation, essentially skipping the error-prone, manual entry of information by the user. For example, PrivacyBee [60] advertises a policy generator on "autopilot," claiming to automatically recognize all privacy-relevant services used on a given website and to generate as well as regularly update the website's policy. For this service, PrivacyBee charges a yearly fee of around $70 per domain.

None of the generators we identify in our study (see Table 5 below) publicly mention the use of generative AI to draft policies. While this may change in the future and our definition does not preclude it, we assume that policy texts are standardized enough for generators to rely entirely on pre-defined clauses that are combined together depending on the user input. Note that we exclude mere templates that users must manually complete or modify from our definition of generators, as generators are based on the notion that the *user is not intended to directly interact with the actual text of the policy*. While we carefully and manually ensure that all generators identified in this study fulfill this definition (see Section 6.2), we acknowledge that it is not always possible to perfectly and objectively distinguish a generator from a template.

6.2 Generator Identification Method

To detect the prevalence of generators, we first sought to identify all major providers. For this purpose, we introduced the exploratory question `org` in our LLM-based analysis: "Does the policy mention that it has been created with the help of a tool (e.g., policy generator) or template?" This very broad definition was used to collect potential generators inclusively, with the aim of discarding false positives in a subsequent manual step. Unlike the nine disclosure dimensions defined in Section 3, the response to `org` is a string value representing the name of the stated candidate generator, and we did not validate the model's output due to its exploratory nature.

Next, we manually inspected the model output to remove false positives. Using the identified generators and associated text in policies, we then created a reliable, conservative dictionary of RegEx patterns that we used to detect acknowledgment of each generator in all policies (see artifact repository). We refined the RegEx patterns until we achieved perfect accuracy on random samples of ten policies per generator. We manually verified that the resulting 140 policies all actually disclosed the relevant generator. As a result, precision is high, and our estimates should be interpreted as a conservative lower bound of generator use. This procedure allowed us to overcome the model's challenge of recognizing *ex ante* whether the stated source of a policy is a generator or a mere template.

6.3 Generator Use and Disclosures

Table 5 presents generator use statistics. Generator use is pervasive, with over 2,000 policies explicitly produced by one in the October dataset. Their use accounts for 18.2% of all policies in the October `CH` group. We observe disproportionately higher use of generators for Swiss-facing websites. We also found use of EU generators for Swiss-facing websites, but *not* the opposite.

In Table 6, we show generator use by policy language and top website rank according to CrUX, i.e., the highest popularity bucket across all countries (see Section 4), for all October policies. In this table and in all the following ones, we analyze the policies of all three groups together. Generators have predominantly been used for German policies (16.4%), with all other languages averaging less than 4%. This is likely due to many generators only being available in German, which may explain the discrepancy in generator use between Swiss- and EU-facing websites, as the latter group contains a smaller relative proportion of German-language policies (see Appendix A.4, Table 11).

Moreover, generators are mostly used by less popular websites, with less than 5% of the policies from Top 5k websites indicating use. This intuitively makes sense: larger websites are more exposed to legal and reputational risks, and they presumably have more resources to hire internal or external legal advisors.

We also inspected how generator use affects the disclosures of policies. Table 7 illustrates that policies created with a generator are associated with increased disclosure rates of up to 15 p.p. (right to lodge a complaint). Disclosure of the controller's identity—required by both Swiss and EU privacy law—is also noticeably higher (+9.5

	EU		CH		CH & EU	
Generator[1]	Aug	Oct	Aug	Oct	Aug	Oct
SwissAnwalt (CH)	0	0	549	539	137	123
Datenschutzpartner (CH)	0	0	104	239	46	98
eRecht24 (DE)	9	9	146	136	11	11
PrivacyBee (CH)	0	0	55	177	20	45
DGD (DE)	28	29	64	61	40	34
activeMind (DE)	21	21	63	62	16	15
BrainBox (CH)	0	0	49	96	16	29
Schwenke (DE)	17	17	43	39	24	23
AdSimple (DE)	9	9	41	43	5	8
WeissPartner (DE)	5	5	32	30	12	12
Iubenda (IT)	5	5	20	23	11	15
MeinDatenschutz (DE)	9	10	16	16	5	5
LegallyOK (CH)	0	0	6	32	1	7
TrustedShops (DE)	0	0	14	3	5	5
With generator[2]	102	104	1200	1492	347	424
% of all policies	4.9	4.8	17.1	18.2	10.3	10.7

[1] Name with country of origin in parenthesis.
[2] Policies stating having used *at least one* generator. Only 13 policies across all groups state having used two or more generators.

Table 5: Prevalence of generators (sorted by total use).

	Language				Top Website Rank		
	DE	EN	IT	FR	5k	10–50k	100k+
Total policies	11855	1600	725	153	1108	7224	6001
With generator	1945	62	13	0	47	1000	973
Percentage	16.4%	3.9%	1.8%	0.0%	4.2%	13.8%	16.2%

Table 6: Prevalence of generators by policy language and website popularity rank (all groups, October 2023).

p.p.). This pattern holds even when disaggregating generator use by website popularity rank (see Appendix A.12).

The only exception is automated decisions, which are mentioned by 7.5 p.p. fewer policies with a generator. This could be due to the fact that generators are mostly used by smaller websites (see Table 6), which perform automated decision-making less frequently than larger websites (Top 5k websites refer to automated decisions 13.2 p.p. more than non-Top 5k websites, see Appendix A.7, Table 15). Without knowing which websites actually rely on automated decisions, we lack a conclusive explanation.

Finally, we inspected the heterogeneous disclosure across all individual generators in our sample and present our findings for the Top 5 generators in Table 8 (see Appendix A.12 for the full list). While most generators achieve higher average disclosure rates than the 67.3% observed for policies without any generator use (Table 7), there are two prominent exceptions: SwissAnwalt and eRecht24.

SwissAnwalt illustrates the impact of choice architecture and default configurations used by generators. As shown in Figure 6, the generator provided an option to include data subject rights in the resulting policy. Since Swiss data protection law does not mandate

	Generator Used		
Obligation	No	Yes	Δ p.p.[1]
contr	78.8%	88.3%	+ 9.5
purp	98.6%	100.0%	+ 1.3
rect	79.7%	83.8%	+ 4.0
forg	81.8%	84.0%	+ 2.1
port	56.0%	69.0%	+13.0
comp	53.7%	68.7%	+15.0
hum	22.6%	15.1%	-7.5
Average	67.3%	72.7%	+ 5.3
Total policies	12313	2020[2]	–

1. The Δ column reports percentage-point differences for generator use.
2. October policies stating having used *at least one* generator.

Table 7: Disclosures by generator use (October 2023).

the disclosure of these rights, the corresponding selection box was *not* pre-checked. Consequently, data subject rights appeared in the generated output only when users explicitly opted in. Although a substantial share of SwissAnwalt users nonetheless chose to disclose these rights, many did not. Given the high disclosure rates of the other two major Swiss generators, Datenschutzpartner and PrivacyBee, we suspect they mentioned these rights by default.

The story with eRecht24 is different: policies generated with the tool simultaneously exhibit very high values (like the right of access and rectification) and very low values (like the right to data portability). Having inspected all October policies referencing eRecht24, we found the following exact boilerplate text in 92.3% of them (translated from German):

> *You have the right to receive information for free at any time about your stored personal data, its origin and recipients, and the purpose of data processing, as well as the right to rectify, block, or delete this data.*

The text does not mention the rights to data portability and to lodge a complaint. Even though eRecht24 has been providing a checklist that states that these two rights must be mentioned in the policy since at least 2022 [68], most websites have apparently not updated their policies accordingly. This is particularly salient given that eRecht24 is based in Germany, where the GDPR requires websites to disclose all data subject rights.

6.4 Standardization and Clusters

The use of standard clauses for disclosures is not limited to eRecht24: the vast majority (88.3%) of policies generated with PrivacyBee—which exhibit very high transparency disclosures—contain the exact same paragraph disclosing user rights. This shows how textual standardization can be either beneficial or detrimental depending on the quality of the standardized clause. It also demonstrates the path dependencies of generated policies if their pre-defined clauses are not updated over time. For the generator Datenschutzpartner, by contrast, we found evidence for several different versions of the transparency disclosures.

Generator	*contr*	*purp*	rect	forg	port	comp	*hum*	**Average**	**Policies**	**Market Share**
SwissAnwalt (CH)	92.5%	100.0%	57.8%	58.3%	45.5%	46.1%	1.4%	57.4%	657	32.7%
Datenschutzpartner (CH)	99.1%	99.7%	99.7%	99.7%	89.3%	99.7%	0.6%	84.0%	337	16.7%
PrivacyBee (CH)	99.5%	100.0%	99.5%	99.5%	98.6%	99.5%	47.7%	92.1%	222	11.0%
eRecht24 (DE)	5.8%	100.0%	98.1%	98.1%	5.8%	5.8%	1.3%	45.0%	156	7.8%
DGD (DE)	98.3%	100.0%	98.3%	98.3%	97.5%	95.0%	99.2%	98.1%	121	6.0%

Table 8: Disclosure rates by Top 5 generator and obligation for policies using a single generator (October 2023).

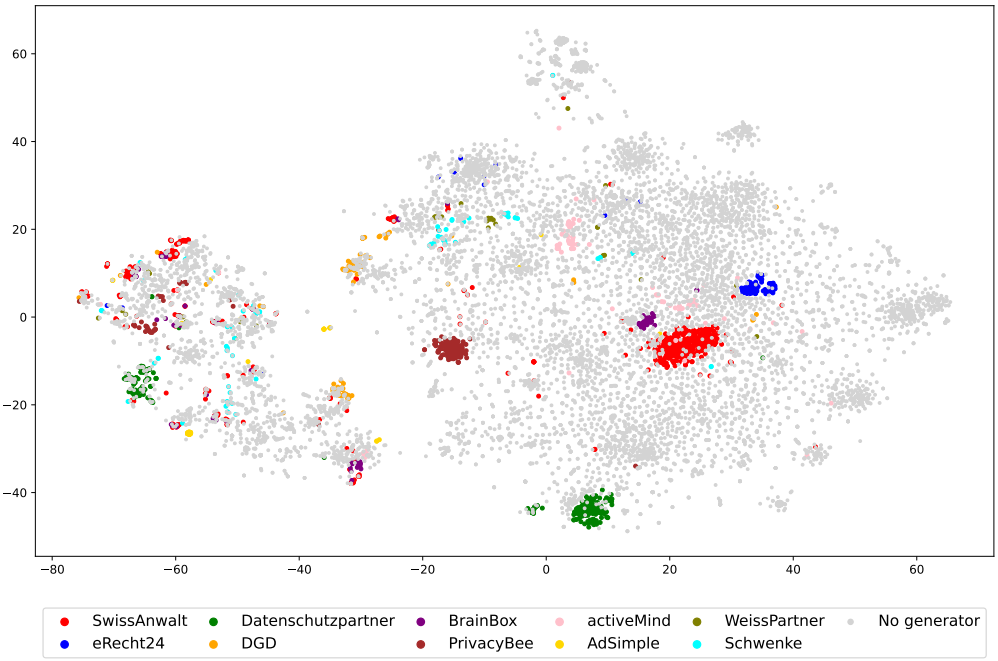


Figure 4: Cluster analysis of policies by generator (Top 10) for October policies in German ($N = 11,855$).

These observations motivated us to systematically analyze the extent to which generators output standardized policies. This matters both in terms of disclosures mandated by Swiss privacy law as well as the diffusion of the GDPR on a voluntary basis, as generators with standardized clauses exhibiting a high market share can significantly affect both dimensions. In short, we find evidence consistent with a relatively high degree of standardization.

We assess standardization in generator-produced policies by embedding all German-language policies from the October snapshot and projecting them to two dimensions using t-SNE [45] based on cosine similarity. Similar visualization approaches have been used to analyze standardization in legal documents [10, 11]. These traditional approaches first used term-document matrices like TF-IDF to vectorize the text data. By contrast, we used the text embedding model `text-embedding-3-large` from OpenAI [55] to preserve the semantic structure of the policies. The embeddings were based on each policy's first 8,192 tokens, due to the model's context window limitation. While some policies exceed this window, Badawi et al. [10, p. 16] demonstrated strong performance clustering legal documents with only the first 1,000 characters. We colored the policies of the ten most widely used generators (see Table 5).

As shown in Figure 4, generators form distinct, albeit imperfect, clusters. One possible reason for these imperfect clusters would be that websites modify the generator's output. For instance, one website stated (translated from German):

> *This privacy policy was created using the privacy policy generator from DGD [...] with a lawyer specializing in data protection law, and the model privacy policy provided by the law firm Weiß & Partner [67].*

The provision of different outputs for different data processing practices is, moreover, an inherent feature of generators. Generated policies that fall outside corresponding clusters may simply originate from websites with unusual data practices (e.g., the collection of atypical data). Furthermore, generators evolve over time: different versions of the same generator (e.g., before and after the GDPR) likely coexist in the October corpus of policies.

7 Discussion

7.1 Assessment and Policy Implications

Our analysis offers evidence that the revised Swiss privacy law significantly affected Swiss-facing websites' policies. Moreover, we find that the revision led to an increase of voluntary disclosures of data subject rights, which would be consistent with—but not direct evidence of—a delayed Brussels Effect of the GDPR in Switzerland. Taken together, these findings highlight that, by revising their privacy laws, smaller jurisdictions can meaningfully increase both mandatory and voluntary privacy disclosures.

Despite the revised Swiss privacy law having entered into force back in 2023, it remains relevant to this very day. The Swiss data protection authority [25] announced having received 1,406 reports of violations of the Swiss privacy law from its entry into force through January 2025. With Switzerland contemplating whether and how to regulate artificial intelligence, both the Swiss Federal Office of Justice [26] and a prominent legal scholar [74] have also proposed to extend the current scope of the right to be informed about automated decision-making (hum).

The generally strong performance of GPT-5 in assessing policy disclosures in our multilingual pipeline raises two noteworthy discussion points. First, our dataset demonstrates how transparency disclosures can be reliably identified in different languages with a single English prompt. This implies that regulators in multilingual jurisdictions can efficiently evaluate policy disclosures in German, French, Italian, and English using our pipeline. The two caveats are (1) that there must be an underlying codebook and annotated dataset, which we release with our study, and (2) that the method does not guarantee perfect performance, particularly beyond the reasonably objective attributes we considered. Nonetheless, we argue that having our method assist regulators in identifying potential non-compliant policies could provide significant benefits. Second, as our performance for the automated decision-making metric hum indicates (see Figure 1), LLMs may still struggle to assess specific dimensions of disclosures, which underlines the importance of both carefully drafted codebooks and cautious AI use.

Finally, we find that generators are not only particularly prevalent for smaller German-language websites, but that their use is connected to higher rates of transparency disclosures in the context of this study. Policymakers should take note, as generators need not be for-profit: the UK's Information Commissioner's Office provides a free policy generator [36]. We welcome this effort, as our findings suggest that generators can support small business compliance. This is not to say that generators are perfect: Pan et al. [56] document flaws in the synthetically generated policies they inspect, and we fear that issues in generated policies may not be subsequently rectified (see Section 6.3). Policymakers might wish to consider generators as an intervention point, encouraging routine auditing as well as timely updates of both the system itself and outdated generated policies.

7.2 Limitations

Our study has limitations typical of empirical work in this area. The short temporal time window was deliberately chosen to avoid weakening our causal identification method by measuring unrelated subsequent events. We concede that the reported changes likely constitute lower bounds of the revision's true effect, as some organizations may have opted to comply at a later time. Our grouping procedure, moreover, inevitably oversimplifies the territorial scope of Swiss and European privacy law, but we believe it approximates the websites subject to each privacy law regime reasonably well. Our dataset also focuses on a restricted set of variables. Additional dimensions such as website categories could be used as controls in the regression analysis.

While manual annotations inevitably involve a degree of subjectivity, we alleviated this issue through the iterative refinement of our codebook until annotators reached at least acceptable agreement. Furthermore, the LLM-based policy analysis cannot guarantee perfect reproducibility, since model outputs may vary slightly across runs. Finally, our conservative generator identification method minimizes the risk of false positives, potentially at the cost of underestimating the true use of generators for drafting policies.

Despite these limitations, our paper sheds light on (1) the real-world effects of a privacy law revision beyond the widely studied European and American privacy laws, (2) the potential of LLMs at large-scale multilingual disclosure assessments, and (3) the important role that generators play in website policies.

8 Conclusion and Future Work

This paper extends our understanding of the real-world impact of privacy law changes by investigating the effects of a Swiss privacy law revision for 35,000 websites. Developing and validating a novel multilingual disclosure assessment pipeline for policies, we demonstrate that the revision significantly increased privacy disclosures required by Swiss or European law. We also find that generators can play an important role in helping smaller websites comply with privacy legislation. This work paves the way for future research along the following dimensions and others:

(1) How do privacy law revisions in different jurisdictions affect transparency disclosures in policies?
(2) Do larger organizations with multilingual or multijurisdictional policies voluntarily implement privacy law revisions across all their policy versions?
(3) To what extent can our multilingual disclosure assessment pipeline be used for different types of legal documents, such as terms of service?
(4) Is generator use equally widespread in other countries and for different legal documents?
(5) What can policymakers draw from these findings to improve the details of privacy laws, monitoring and enforcement, and quality and ease of transparency disclosures?

Pursuing this line of research could yield three key benefits for enhancing privacy: (1) assessing the efficacy of privacy laws from different jurisdictions can help us understand which approaches work and which do not; (2) LLMs may support regulators in monitoring specific aspects of compliance at scale; and (3) generators may assist with privacy-related disclosures, particularly for resource-constrained smaller businesses. We hope our work will spark future research in these directions.

Acknowledgments

The authors used generative AI-based tools to revise the text, improve flow, and correct any typos, grammatical errors, and awkward phrasing. The authors also acknowledge having used AI-based tools for assistance in analyzing the data and editing LaTeX.

The authors would like to thank Elias Landes for his excellent research assistance.

Luka Nenadic gratefully acknowledges the support of the Swiss National Science Foundation (project 10002634).

The work of David Rodriguez was partially supported by the Re-InitS project, funded by the Plan Estatal de Investigación Científica y Técnica y de Innovación 2024–2027 (Ministerio de Ciencia e Investigación (Spain) - MCIN/AEI/10.13039/501100011033) under grant agreement PID2024-155230OB-C43 and by "DINA-iOS: Desarrollo de una infraestructura experimental para la auditoría de aplicaciones iOS" project funded by ETSI Telecomunicación under "Ayudas Primeros Proyectos" grant.

References

[1] Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics* (Online) *(Findings of EMNLP '20)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 743–749. https://doi.org/10.18653/v1/2020.findings-emnlp.66

[2] Chunrong Ai and Edward C. Norton. 2024. Difference in differences, ratio in ratios, and ratio in odds ratios for limited dependent variables: A review and more. *Studies in Nonlinear Dynamics & Econometrics* 28, 4 (Aug. 2024), 1–32. https://doi.org/10.1515/snde-2024-0125

[3] Jose M. Del Alamo, Danny S. Guaman, Boni García, and Ana Diez. 2022. A systematic mapping study on automated analysis of privacy policies. *Computing* 104, 9 (May 2022), 2053–2076. https://doi.org/10.1007/s00607-022-01076-3

[4] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, Stroudsburg, PA, USA, 2165–2176. https://doi.org/10.1145/3442381.3450048

[5] Benjamin Andow, Samin Yaseer Mahmud, William Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating internal privacy policy contradictions on Google Play. In *Proceedings of the 28th USENIX Security Symposium* (Santa Clara, CA, United States) *(USENIX Security '19)*. USENIX Association, Berkeley, CA, USA, 585–602. https://www.usenix.org/conference/usenixsecurity19/presentation/andow

[6] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with PoliCheck. In *Proceedings of the 29th USENIX Security Symposium* (online) *(USENIX Security '20)*. USENIX Association, Berkeley, CA, USA, 985–1002. https://www.usenix.org/system/files/sec20-andow.pdf

[7] Aly Apacible-Bernardo and Kayla Bushey. 2025. Data protection and privacy laws now in effect in 144 countries. Retrieved 2026-02-23 from https://iapp.org/news/a/data-protection-and-privacy-laws-now-in-effect-in-144-countries

[8] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. 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* (Marseille, France) *(LREC '22)*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 5460–5472. Retrieved 2024-05-10 from https://aclanthology.org/2022.lrec-1.585

[9] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (Dec. 2008), 555–596. https://doi.org/10.1162/coli.07-034-r2

[10] Adam B. Badawi, Elisabeth de Fontenay, and Julian Nyarko. 2023. The value of M&A drafting. SSRN Working Paper:4337075 https://ssrn.com/abstract=4337075

[11] Robert Bartlett. 2026. Standardization and innovation in venture capital contracting: Evidence from startup company charters. *The Journal of Legal Studies* 55, 1 (2026), 83–128. https://doi.org/10.1086/736135

[12] Benjamin Barton and Deborah Rhode. 2019. Access to justice and routine legal services: New technologies meet bar regulators. *Hastings Law Journal* 70 (2019), 955–989. Retrieved 2025-09-16 from https://repository.uclawsf.edu/hastings_law_journal/vol70/iss4/2/

[13] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[14] David Bernhard, Luka Nenadic, Stefan Bechtold, and Karel Kubicek. 2025. Multilingual scraper of privacy policies and terms of service. In *Proceedings of the 2025 Symposium on Computer Science and Law* (Munich, Germany) *(CSLAW '25)*. Association for Computing Machinery, New York, NY, USA, 55–63. https://doi.org/10.1145/3709025.3712215

[15] Kathryn D. Betts and Kyle R. Jaep. 2017. The dawn of fully automated contract drafting: Machine learning breathes new life into a decades-old promise. *Duke Law & Technology Review* 15 (2017), 216–233. Retrieved 2025-09-16 from https://scholarship.law.duke.edu/dltr/vol15/iss1/11

[16] Anu Bradford. 2020. *The Brussels Effect.* Oxford University Press, Oxford, United Kingdom.

[17] Chrome. 2024. Overview of CrUX. Retrieved 2025-01-23 from https://developer.chrome.com/docs/crux

[18] Jorge L. Contreras. 2025. Solving for agreement. SSRN Working Paper:5389732 https://ssrn.com/abstract=5389732

[19] Thomas Cory, Wolf Rieder, Julia Krämer, Philip Raschke, Patrick Herbke, and Axel Küpper. 2026. Word-level annotation of GDPR transparency compliance in privacy policies using large language models. *Proceedings on Privacy Enhancing Technologies* 2026, 1 (2026), 509–528. https://doi.org/10.56553/popets-2026-0026

[20] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness: (short paper). In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society* (Raleigh, North Carolina, USA) *(WPES '12)*. Association for Computing Machinery, New York, NY, USA, 91–96. https://doi.org/10.1145/2381966.2381979

[21] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. 2019. Measuring cookies and web privacy in a post-GDPR world. In *Passive and Active Measurement (PAM '19)*, David Choffnes and Marinho Barcellos (Eds.). Springer, Cham, 258–270. https://doi.org/10.1007/978-3-030-15986-3-17

[22] Harshil Darji, Stefan Becher, Jelena Mitrović, Armin Gerl, and Michael Granitzer. 2024. A dataset of GDPR compliant NER for privacy policies. In *Proceedings of the 6th International Open Search Symposium (ossym '24)*. CERN, Garching (Munich), Germany, 26–31. https://doi.org/10.5281/zenodo.13871889

[23] Datenschutzpartner. 2026. Datenschutz-Generator für Datenschutzerklärungen. Retrieved 2026-02-20 from https://www.datenschutzpartner.ch/angebot-datenschutz-generator/

[24] Kevin E. Davis and Florencia Marotta-Wurgler. 2024. Filling the void: How E.U. privacy law spills over to the U.S. *Journal of Law and Empirical Analysis* 1 (June 2024), 1–21. https://doi.org/10.1177/2755323X241237619

[25] Federal Data Protection and Information Commissioner. 2025. Data Protection Day 2025: Data protection in the face of digital transformation. https://www.edoeb.admin.ch/en/data-protection-day-2025

[26] Federal Office of Justice. 2024. *Rechtliche Basisanalyse im Rahmen der Auslegeordnung zu den Regulierungsansätzen im Bereich künstliche Intelligenz.* Technical Report. https://www.bj.admin.ch/bj/de/home/staat/gesetzgebung/kuenstliche-intelligenz.html

[27] William Foster and Andrew Lawson. 2018. When to praise the machine: The promise and perils of automated transactional drafting. *South Carolina Law Review* 69, 3 (April 2018), 597–635. Retrieved 2026-01-27 from https://scholarcommons.sc.edu/sclr/vol69/iss3/5

[28] Jens Frankenreiter. 2022. Cost-based California Effects. *Yale Journal on Regulation* 39 (2022), 1155–1217. Retrieved 2025-05-25 from https://www.yalejreg.com/print/cost-based-california-effects/

[29] Arda Goknil, Frank B. Gelderblom, Sondre Tverdal, Shreyas Tokas, and Heejin Song. 2024. Privacy policy analysis through prompt engineering for LLMs (PAPEL). arXiv:2409.14879 [cs.CL] https://arxiv.org/abs/2409.14879

[30] Google. 2022. Compact Language Detector v3 (CLD3). Retrieved 2025-09-11 from https://github.com/google/cld3

[31] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press, Princeton, NJ, USA.

[32] Danny S. Guaman, David Rodriguez, Jose M. Del Alamo, and Jose Such. 2023. Automated GDPR compliance assessment for cross-border personal data transfers in Android applications. *Computers & Security* 130, Article 103262 (July 2023), 16 pages. https://doi.org/10.1016/j.cose.2023.103262

[33] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam

Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LEGALBENCH: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NeurIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1915, 157 pages.

[34] 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 *Proceedings of the 27th USENIX Security Symposium* (Baltimore, MD, USA) *(USENIX Security '18)*. USENIX Association, Berkeley, CA, USA, 531–548. Retrieved 2024-05-24 from https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[35] Henry Hosseini, Christine Utz, Martin Degeling, and Thomas Hupperich. 2024. A bilingual longitudinal analysis of privacy policies measuring the impacts of the GDPR and the CCPA/CPRA. *Proceedings on Privacy Enhancing Technologies* 2024, 2 (2024), 434–463. https://doi.org/10.56553/popets-2024-0058

[36] Information Commissioner's Office (ICO). 2024. Create your own privacy notice. Retrieved 2025-01-31 from https://ico.org.uk/for-organisations/advice-for-small-organisations/create-your-own-privacy-notice/

[37] Ani Janelidze. 2022. *5 best practices for writing annotation guidelines*. V7 Labs. Retrieved 2025-09-08 from https://www.v7labs.com/blog/annotation-guidelines

[38] Marcel Kahan and Michael Klausner. 1997. Standardization and innovation in corporate contracting (or "the economics of boilerplate"). *Virginia Law Review* 83, 4 (May 1997), 713–770. https://doi.org/10.2307/1073747

[39] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research* 30, 3 (2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[40] Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Bangkok, Thailand) *(ACL '24)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 15339–15353. https://doi.org/10.18653/v1/2024.acl-long.818

[41] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 47–64. https://doi.org/10.2478/popets-2020-0004

[42] Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016. Modeling language vagueness in privacy policies using deep neural networks. In *AAAI 2016 Fall Symposium on Privacy and Language Technologies* (Arlington, VA, USA). Association for the Advancement of Artificial Intelligence, Washington, DC, USA, 257–263. https://aaai.org/papers/14059-14059-modeling-language-vagueness-in-privacy-policies-using-deep-neural-networks/

[43] Fei Liu, Rohan Ramanath, Norman M. Sadeh, and Noah A. Smith. 2014. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of COLING 2014* (Dublin, Ireland). Association for Computational Linguistics, Stroudsburg, PA, USA, 884–894. https://aclanthology.org/C14-1084.pdf

[44] Shuang Liu, Fan Zhang, Baiyang Zhao, Renjie Guo, Tao Chen, and Meishan Zhang. 2023. APPCorp: A corpus for Android privacy policy document structure analysis. *Frontiers of Computer Science* 17, 3 (2023), 173320. https://doi.org/10.1007/s11704-022-1627-2

[45] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605. Retrieved 2025-09-15 from http://jmlr.org/papers/v9/vandermaaten08a.html

[46] Florencia Marotta-Wurgler and David Stein. 2025. Building a long text privacy policy corpus with multi-class labels. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vienna, Austria) *(ACL '25)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 8156–8219. https://doi.org/10.18653/v1/2025.acl-long.401

[47] Malak Mashaabi, Ghadi Al-Yahya, Raghad Alnashwan, and Hend Al-Khalifa. 2023. Arabic privacy policy corpus and classification. In *Natural Language Processing and Information Systems: 28th International Conference on Applications of Natural Language to Information Systems (NLDB 2023), Derby, UK, June 21–23, 2023, Proceedings*. Springer, Cham, 94–108. https://doi.org/10.1007/978-3-031-35320-8-7

[48] Aaron K. Massey, Jacob Eisenstein, Annie I. Antón, and Peter P. Swire. 2013. Automated text mining for requirements analysis of policy documents. In *Proceedings of the 21st IEEE International Requirements Engineering Conference* (Rio de Janeiro, Brazil) *(RE '13)*. IEEE, New York, NY, USA, 4–13. https://www.cc.gatech.edu/~aianton/assets/2013-re13-nlp.pdf

[49] Abraham Mhaidli, Selin Fidan, An Doan, Gina Herakovic, Mukund Srinath, Lee Matheson, Shomir Wilson, and Florian Schaub. 2023. Researchers' experiences in analyzing privacy policies: Challenges and opportunities. *Proceedings on Privacy Enhancing Technologies* 2023, 4 (2023), 287–305. https://doi.org/10.56553/popets-2023-0111

[50] Microsoft. 2025. Presidio: Data protection and de-identification SDK. Retrieved 2025-09-26 from https://microsoft.github.io/presidio/

[51] Maaz Bin Musa, Steven M. Winston, Garrison Allen, Jacob Schiller, Kevin Moore, Sean Quick, Johnathan Melvin, Padmini Srinivasan, Mihailis E. Diamantis, and Rishab Nithyanand. 2024. C3PA: An open dataset of expert-annotated and regulation-aware privacy policies to enable scalable regulatory compliance audits. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Miami, Florida, USA) *(EMNLP '24)*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 3710–3722. https://doi.org/10.18653/v1/2024.emnlp-main.217

[52] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark) *(EMNLP '17)*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 2774–2779. https://doi.org/10.18653/v1/D17-1294

[53] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2018. PrivOnto: A semantic framework for the analysis of privacy policies. *Semantic Web* 9, 2 (Jan. 2018), 185–203. https://doi.org/10.3233/SW-170283

[54] OpenAI. 2024. Introducing structured outputs in the API. Retrieved 2025-09-08 from https://openai.com/index/introducing-structured-outputs-in-the-api/

[55] OpenAI. 2024. New embedding models and API updates. Retrieved 2025-09-15 from https://openai.com/index/new-embedding-models-and-api-updates/

[56] Shidong Pan, Dawen Zhang, Mark Staples, Zhenchang Xing, Jieshan Chen, Xiwei Xu, and Thong Hoang. 2024. Is it a trap? A large-scale empirical study and comprehensive assessment of online automated privacy policy generators for mobile apps. In *Proceedings of the 33rd USENIX Conference on Security Symposium* (Philadelphia, PA) *(USENIX Security '24)*. USENIX Association, Berkeley, CA, USA, 5681–5698. https://doi.org/10.5555/3698900.3699218

[57] Harshvardhan J. Pandit, Beatriz Esteves, Georg P. Krog, Paul Ryan, Delaram Golpayegani, and Julian Flake. 2024. Data Privacy Vocabulary (DPV) – Version 2. arXiv:2404.13426 [cs.CY] https://arxiv.org/abs/2404.13426

[58] Przemysław Pałka, Radosław Pałosz, and Katarzyna Wiśniewska. 2023. *Annotated privacy policies of 100 online platforms*. https://doi.org/10.17632/pcgvm6zh43.1

[59] Christian Peukert, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. 2022. Regulatory spillovers and data governance: Evidence from the GDPR. *Marketing Science* 41, 4 (Feb. 2022), 746–768. https://doi.org/10.1287/mksc.2021.1339

[60] PrivacyBee. 2025. Datenschutz für deine Webseite auf Autopilot. Retrieved 2025-09-12 from https://www.privacybee.io/ch/

[61] Abhilasha Ravichander, Alan W. Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (Hong Kong, China) *(EMNLP-IJCNLP '19)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 4947–4958. https://doi.org/10.18653/v1/D19-1500

[62] David Rodriguez, Jose M. Del Alamo, Celia Fernández-Aller, and Norman Sadeh. 2024. Sharing is not always caring: Delving into personal data transfer compliance in Android apps. *IEEE Access* 12 (2024), 5256–5269. https://doi.org/10.1109/ACCESS.2024.3349425

[63] David Rodriguez, Celia Fernández-Aller, Jose M. Del Alamo, and Norman Sadeh. 2024. Data retention disclosures in the Google Play Store: Opacity remains the norm. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW '24)*. IEEE, Vienna, Austria, 19–23. https://doi.org/10.1109/EuroSPW61312.2024.00009

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

[65] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. 2016. Automatic extraction of opt-out choices from privacy policies. In *AAAI 2016 Fall Symposium on Privacy and Language Technologies* (Arlington, VA, USA). Association for the Advancement of Artificial Intelligence, Washington, DC, USA, 270–275. https://aaai.org/papers/14114-14114-automatic-extraction-of-opt-out-choices-from-privacy-policies/

[66] Charles S. Saxon. 1982. Computer-aided drafting of legal documents. *American Bar Foundation Research Journal* 7, 3 (July 1982), 685–754. https://doi.org/10.1111/j.1747-4469.1982.tb00469.x

[67] Schneemenschen GmbH. 2018. Datenschutzerklärung. Retrieved 2025-09-12 from https://perma.cc/E9W9-6U9Z

[68] Sören Siebert. 2025. DSGVO-konforme Datenschutzerklärung jetzt kostenlos erstellen. Retrieved 2025-09-26 from https://www.e-recht24.de/muster-datenschutzerklaerung.html

[69] Martin Steiger. 2022. Deutscher Datenschutz-Generator: 250 Euro-Abmahnung für eine Datenschutzerklärung? Retrieved 2025-09-12 from https://steigerlegal.ch/2022/06/25/datenschutz-generator-abmahnungen-deutschland/

[70] Ruoxi Sun and Minhui Xue. 2020. Quality assessment of online automated privacy policy generators: An empirical study. In *Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering* (Trondheim, Norway) *(EASE '20)*. Association for Computing Machinery, New York, NY, USA, 270–275. https://doi.org/10.1145/3383219.3383247

[71] Swiss Confederation. 2020. Federal Act on Data Protection (Data Protection Act). Retrieved 2024-05-10 from https://www.fedlex.admin.ch/eli/cc/2022/491/en

[72] SwissAnwalt. 2023. Datenschutz Generator für Webseiten, Blogs und Social Media - Kostenlos. Retrieved 2025-09-12 from https://web.archive.org/web/20230310013530/https://www.swissanwalt.ch/datenschutz-generator.aspx

[73] Chenhao Tang, Zihan Liu, Chuxin Ma, Zihao Wu, Yuxuan Li, Weiyu Liu, Dongruo Zhu, Qizhang Li, Xiang Li, Ting Liu, and Lei Fan. 2023. PolicyGPT: Automated analysis of privacy policies with large language models. arXiv:2309.10238 [cs.CL] https://arxiv.org/abs/2309.10238

[74] Florent Thouvenin. 2026. Ein Rechtsrahmen für KI in der Schweiz: Handlungsbedarf und Handlungsoptionen bei Transparenz und Datenschutz. *Jusletter* March 30, 2026 (2026), 1–21. https://jusletter.weblaw.ch/juslissues/2026/1278/ein-rechtsrahmen-fur_995089b415.html__ONCE&login=false

[75] Van Hong Tran, Aarushi Mehrotra, Marshini Chetty, Nick Feamster, Jens Frankenreiter, and Lior Strahilevitz. 2024. Measuring compliance with the California Consumer Privacy Act over space and time. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, 1–19. https://doi.org/10.1145/3613904.3642597

[76] Tina Tseng, Amanda Stent, and Domenic Maida. 2020. Best practices for managing data annotation projects. arXiv:2009.11654 [cs.CY] Retrieved 2025-09-08 from https://arxiv.org/abs/2009.11654

[77] Isabel Wagner. 2023. Privacy policies across the ages: Content of privacy policies 1996–2021. *ACM Transactions on Privacy and Security* 26, 3, Article 32 (May 2023), 32 pages. https://doi.org/10.1145/3590152

[78] 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* (Berlin, Germany) *(ACL '16)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1330–1340. https://doi.org/10.18653/v1/P16-1126

[79] Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability – An empirical approach to inter-rater reliability. 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)* (Online) *(ACL-IJCNLP '21)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 7053–7065. https://doi.org/10.18653/v1/2021.acl-long.548

[80] Qinge Xie, Karthik Ramakrishnan, and Frank Li. 2025. Evaluating privacy policies under modern privacy laws at scale: An LLM-based automated approach. In *Proceedings of the 34th USENIX Security Symposium* (Seattle, WA, USA) *(USENIX Security '25)*. USENIX Association, Berkeley, CA, USA, 5797–5816. https://www.usenix.org/conference/usenixsecurity25/presentation/xie

[81] Ming Yang, Vijayalakshmi Atluri, Shamik Sural, and Atul Kundu. 2025. Automated privacy policy analysis using large language models. In *Data and Applications Security and Privacy XXXIX* (Gjovik, Norway) *(DBSec '25)*. Springer, Cham, Switzerland, 23–43. https://doi.org/10.1007/978-3-031-96590-6-2

[82] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R. Reidenberg, N. Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 66–86. https://doi.org/10.2478/popets-2019-0037

[83] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated analysis of privacy requirements for mobile apps. In *Proceedings 2017 Network and Distributed System Security Symposium*. Internet Society, San Diego, CA, 1–15. https://doi.org/10.14722/ndss.2017.23034

[84] Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, Mete Sertkan, and Allan Hanbury. 2020. DEXA: Supporting non-expert annotators with dynamic examples from experts. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2109–2112. https://doi.org/10.1145/3397271.3401334

A Appendix

A.1 Scraper Performance

We report the detailed scraper performance metrics as drawn and adapted from Bernhard et al. [14, p. 59] in Table 9.

	English	German	French	Italian	**Average**
Accuracy	0.73	0.70	0.75	0.72	0.73
Recall	0.88	0.78	0.81	0.91	0.85
Precision	0.78	0.86	0.77	0.73	0.79
F_1 score	0.83	0.81	0.79	0.81	0.81

Table 9: Scraper's policy classification metrics.

A.2 Bootstrap Confidence Intervals

Language	Practice	F_1	CI Low	CI High	Positives (GT)
English	ispol	0.933	0.837	1.000	22
English	contr	0.929	0.812	1.000	13
English	purp	0.933	0.842	1.000	22
English	rect	0.971	0.903	1.000	17
English	forg	1.000	1.000	1.000	17
English	port	0.909	0.727	1.000	11
English	comp	1.000	1.000	1.000	9
English	hum	0.600	0.000	0.889	3
German	ispol	0.983	0.947	1.000	29
German	contr	0.909	0.786	1.000	15
German	purp	0.983	0.947	1.000	29
German	rect	0.950	0.867	1.000	19
German	forg	0.977	0.917	1.000	21
German	port	0.966	0.880	1.000	14
German	comp	1.000	1.000	1.000	12
German	hum	0.727	0.286	1.000	4
French	ispol	1.000	1.000	1.000	27
French	contr	0.833	0.686	0.941	15
French	purp	1.000	1.000	1.000	27
French	rect	0.927	0.827	1.000	19
French	forg	1.000	1.000	1.000	20
French	port	0.941	0.769	1.000	8
French	comp	1.000	1.000	1.000	10
French	hum	0.727	0.333	1.000	4
Italian	ispol	1.000	1.000	1.000	22
Italian	contr	0.950	0.869	1.000	20
Italian	purp	1.000	1.000	1.000	22
Italian	rect	0.977	0.919	1.000	21
Italian	forg	1.000	1.000	1.000	21
Italian	port	0.933	0.815	1.000	15
Italian	comp	1.000	1.000	1.000	14
Italian	hum	0.333	0.000	0.800	4
Overall	ispol	0.980	0.957	0.995	100
Overall	contr	0.905	0.849	0.951	63
Overall	purp	0.980	0.959	0.995	100
Overall	rect	0.956	0.918	0.983	76
Overall	forg	0.994	0.980	1.000	79
Overall	port	0.939	0.878	0.981	48
Overall	comp	1.000	1.000	1.000	45
Overall	hum	0.632	0.438	0.788	15

Notes. 95% confidence intervals (CI) are computed via policy-level bootstrap with replacement (1,000 resamples; fixed seed). "Positives (GT)" reports the number of positive cases per practice and language.

Table 10: Validation performance (F_1) with 95% bootstrap confidence intervals by language and practice.

To quantify the statistical stability of our validation results, we compute non-parametric 95% confidence intervals for all F_1 scores using policy-level bootstrap resampling (1,000 iterations; fixed random seed). Resampling is performed at the document level to preserve within-policy dependence across disclosure dimensions.

Table 10 reports point estimates together with bootstrap intervals and the number of positive cases annotated per obligation and language. For most disclosure dimensions, the intervals remain narrow across languages, indicating stable performance estimates. Wider intervals arise for hum, which reflects the conditional nature of this obligation and its lower prevalence in the benchmark.

Overall, the bootstrap analysis corroborates the robustness of the multilingual validation results while transparently characterizing residual uncertainty across obligations.

A.3 Last Updated Date (October Policies)

Looking at the last updated dates (upd) for October policies, Figure 5 shows that most policies with an explicitly mentioned date ($N = 5,867$) were last updated in 2023: either before the Swiss privacy law revision (29.2%) or shortly thereafter (20.3%).

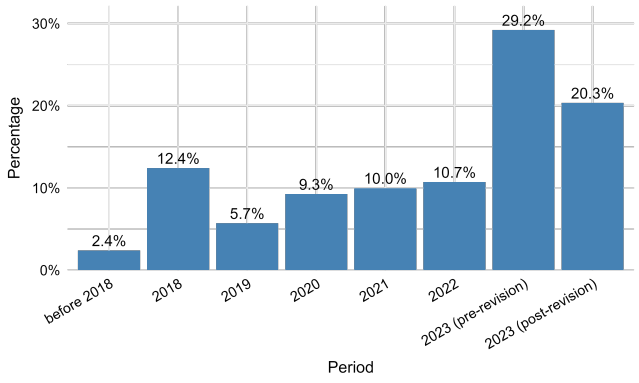


Figure 5: Last updated dates as a percentage of all October policies that are explicitly dated ($N = 5,867$).

A.4 Policy Language Distribution

We present the full group-level language distributions in Table 11.

	EU		CH		CH & EU	
	Aug	Oct	Aug	Oct	Aug	Oct
Websites	5903	5903	20477	20477	9093	9093
% With policy	35.25	36.86	34.19	40.02	37.12	43.57
% in German	57.95	56.94	90.55	90.73	79.11	80.29
% in English	17.64	18.75	7.34	7.08	16.41	15.45
% in French	1.44	1.10	0.80	0.90	1.45	1.39
% in Italian	22.97	23.21	1.31	1.29	3.02	2.88

Table 11: Dataset summary statistics.

A.5 Policy Length and Readability Metrics

Table 12 shows the mean and median word counts for all groups.

	EU	CH	CH & EU
# Words August	2243	1818	2341
# Words October	2257	2125	2643
Δ # words	+14	+307	+302
Δ percentage	+0.6%	+16.9%	+12.9%

Table 12: Median policy word counts.

A.6 GDPR and FADP Terms

We compiled two multilingual lists of terms associated with the GDPR (Table 13) and the FADP (Table 14) to create Table 2.

GDPR Terms
GDPR
2016/679
General Data Protection Regulation
DSGVO
DS-GVO
Datenschutz-Grundverordnung
Datenschutzgrundverordnung
Datenschutz Grundverordnung
règlement général sur la protection des données
RGPD
regolamento generale sulla protezione dei dati

Table 13: GDPR mention terms.

FADP Terms
235.1
Federal Act on Data Protection
Data Protection Act
DSG
nDSG
rDSG
Bundesgesetz über den Datenschutz
Datenschutzgesetz
LPD
Loi fédérale sur la protection des données
Legge federale sulla protezione dei dati

Table 14: FADP mention terms.

A.7 Disclosure Heterogeneity

Table 15 disaggregates October disclosure rates by update status (between the two snapshots), website popularity, and language across

	October Update			Top 5k Website			Language			
Obligation	no	yes	Δ p.p.[1]	no	yes	Δ p.p.[1]	de	en	it	fr
contr	75.0%	88.8%	+13.9	79.5%	86.8%	+7.3	79.7%	79.9%	88.7%	75.2%
purp	98.8%	98.8%	−0.0	98.9%	98.1%	−0.8	98.9%	98.0%	98.5%	99.3%
rect	77.2%	85.6%	+8.4	79.9%	85.1%	+5.2	81.2%	75.1%	80.1%	64.1%
forg	79.5%	86.6%	+7.1	81.8%	86.7%	+5.0	82.9%	77.9%	81.7%	68.6%
port	53.2%	65.9%	+12.7	57.2%	66.2%	+9.0	58.4%	54.9%	60.4%	32.0%
comp	51.7%	62.8%	+11.1	54.8%	67.6%	+12.8	56.7%	49.4%	60.6%	26.8%
hum	19.3%	25.4%	+6.2	20.5%	33.8%	+13.2	21.2%	24.9%	23.6%	7.2%
Average	64.9%	73.4%	+8.5	67.5%	74.9%	+7.4	68.4%	65.7%	70.5%	53.3%
Total policies	9044	5289	–	13225	1108	–	11855	1600	725	153

1. The Δ columns report percentage-point differences for the *yes* subsets versus the *no* subsets.

Table 15: Disclosure per obligation across updated status, website rank, and languages (October 2023).

October policy status	Average disclosure rate	Policies
Unchanged text	64.9%	9,044
Newly valid in October	70.4%	2,419
Changed text	76.0%	2,870

Table 16: Average disclosure rates by October policy status.

all groups. The table provides the obligation-level results underlying the summary reported in Section 5.2. Policies that were either newly valid in October or textually changed between August and October display substantially higher disclosure rates than unchanged policies (+8.5 p.p.). Table 16 further separates this pattern into policies whose text changed between snapshots ("Changed text"), policies that became valid only in October ("Newly valid in October"), and unchanged policies ("Unchanged text"). This decomposition is consistent with active revision during the observation window, though it may also reflect broader standardization and learning dynamics [38] in policy drafting. The descriptive decomposition, however, does not allow these mechanisms to be disentangled.

Website popularity is also associated with higher disclosure rates. Websites included in any country-specific Top 5k bucket exhibit, on average, +7.4 p.p. higher disclosure rates than less visited websites (Table 15). Language-related variation is comparatively modest. Differences between German- and English-language policies remain below 3 p.p. on average. Italian-language policies exhibit comparatively high disclosure levels, and French-language policies display lower overall rates. The smaller sample sizes for Italian ($N = 725$) and French ($N = 153$), however, warrant cautious interpretation.

Finally, Table 19 reports interaction models using an aggregate disclosure index to assess whether the differential August–October increases are concentrated among particular subsets of websites. The results indicate that website popularity is primarily associated with disclosure levels rather than with the reform-related increase itself. By contrast, the increase is strongest among German-language policies, while the non-German interaction is negative, particularly for the CH group.

A.8 Manual Annotation Corrections

In Table 17, we present all changes performed to the annotations by our legal expert, including the relevant snippet for the change. Naturally, the legal expert could not identify the relevant snippet if they did not detect the relevant disclosure (see Policy 93). Policy 77 (denoted with a star *) represents a special case, as it constitutes a website imprint that was erroneously annotated as a valid policy, even though it did not mention personal data at all.

A.9 Impact of Annotation Corrections on Model Performance

To ensure transparency regarding the post-annotation corrections described in Appendix A.8, we additionally evaluated GPT-5 on the original annotations prior to the corrections performed by the legal expert. The original annotations are publicly available in our artifact repository.

Overall, performance differences between the original and corrected annotations are small. Across all language–practice combinations, the mean change in F_1 is 0.021, with a median of 0.000. Table 18 reports the differences in F_1 scores between the corrected and original annotations across languages and practices. The largest improvements occur in dimensions with very low numbers of positive cases, particularly hum, where even a single annotation change can substantially affect the resulting F_1 score. For the remaining dimensions, differences are minor and do not alter the substantive conclusions reported in Section 3.3. No systematic performance regressions are observed across languages or practices.

These findings indicate that the corrections primarily improved annotation accuracy without materially affecting the evaluation outcomes. Importantly, the corrections were identified through manual review of disagreement cases between model outputs and human annotations, and were applied only where the policy text unambiguously supported one interpretation. The robustness of model performance across both annotation versions supports the validity of the reported results.

Entry	Original	Correction	Relevant Snippet
Policy 2: hum	0	1	Art. 22 GDPR: In cases of solely automated processing with legal effects, special rights are granted for data subjects.
Policy 5: upd	24/09/2024	24/09/2018	Version: #2018-09-24
Policy 6: purp	0	1	Personal information submitted to us through our website will be used for the purposes specified in this policy. [...]
Policy 21: purp	0	1	The personal data entered by the data subject is collected and stored solely for use at the location of the data controller and its own purposes.
Policy 23: rect	0	1	You may have further statutory rights, such as the right to have data erased or corrected, to have data processing restricted and to have data transmitted.
Policy 67: port	0	1	A certaines conditions, vous pouvez nous demander de vous transmettre à vous ou à un tiers désigné par vous, vos données personnelles dans un format courant.
Policy 72: contr	0	1	Identité du responsable de traitement : Les données personnelles sont collectées par [...]
Policy 77: contr	1	0	NA*
Policy 77: purp	1	0	NA*
Policy 77: rect	1	0	NA*
Policy 77: forg	1	0	NA*
Policy 77: port	1	0	NA*
Policy 77: comp	1	0	NA*
Policy 77: hum	1	0	NA*
Policy 81: contr	0	1	Pour toute question ou demande liée au traitement de vos données personnelles, vous pouvez nous contacter par e-mail à <redacted email> ou par courrier postal à l'adresse : [...]
Policy 93: contr	1	0	NA
Policy 97: purp	0	1	In particolare, la scuola, ove pubblichi foto di alunni e insegnanti, tratterà delle immagini e lo farà al solo fine di documentare attività didattiche di particolare pregio avendo cura di non rendere identificabili gli alunni in mancanza di base giuridica che non si identifichi nel perseguimento di un interesse pubblico.
Policy 101: contr	0	1	Responsabile per la protezione dei dati: [...]
Policy 112: upd	NA	25/05/2018	Questa policy, in vigore dal 25 maggio 2018 [...]
Policy 113: upd	NA	25/05/2018	La presente Privacy Policy è aggiornata alla data del 25 maggio 2018
Policy 114: upd	01/07/2019	01/06/2019	Ultimo aggiornamento: giugno 2019

Table 17: Overview of the detected errors between the originally annotated and the corrected benchmark.

Lang.	comp	contr	forg	hum	ispol	port	purp	rect
EN	0.000	0.000	0.000	0.156	0.000	0.000	0.050	0.030
FR	0.048	0.091	0.024	0.061	0.000	0.118	0.018	0.022
DE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IT	0.000	0.050	0.000	0.000	0.000	0.000	0.023	0.000
Overall	0.011	0.038	0.006	0.053	0.000	0.020	0.020	0.013

Table 18: Differences in F_1 scores (ΔF_1) between evaluations using the corrected and original annotations. Positive values indicate improved performance after correction.

A.10 Logistic Difference-in-Differences Specification

To assess the sensitivity of the results to functional form, we re-estimate the difference-in-differences models using logistic regression. This nonlinear specification models disclosure on the log-odds scale and constrains fitted values to the unit interval.

All specifications match the analysis in Section 5. For each disclosure obligation, we estimate a logistic regression on the same balanced panel of $N = 11{,}800$ websites observed in both August and October 2023 (23,600 policy–snapshot observations). The dependent variable indicates whether the obligation is disclosed. The model includes indicators for time (October vs. August), group membership (EU as the reference), and their interaction terms. As in the linear models, we include CrUX rank categories and policy language as controls and cluster standard errors at the website level.

Moderator	Term	Coef. (p.p.)	95% CI
Top 5k website	`post × CH`	+2.41	[2.07, 2.75]
	`post × CH & EU`	+2.32	[1.80, 2.85]
	`post × CH × Top 5k`	-2.46	[-5.05, 0.13]
	`post × CH & EU × Top 5k`	-0.71	[-2.20, 0.78]
Non-German policy	`post × CH`	+2.51	[2.13, 2.90]
	`post × CH & EU`	+2.40	[1.84, 2.97]
	`post × CH × Non-German`	-2.18	[-2.98, -1.39]
	`post × CH & EU × Non-German`	-0.90	[-2.01, 0.20]

OLS linear probability models estimated on the balanced panel of 11,800 websites (23,600 policy–snapshot observations). Standard errors are clustered at the website level.

Table 19: Heterogeneity interaction models using the aggregate disclosure index.

Obligation	CH: OR (95% CI)	CH & EU: OR (95% CI)
contr	1.10* (1.07–1.14)	1.11* (1.06–1.16)
purp	0.84 (0.67–1.06)	0.80 (0.63–1.01)
rect	1.10* (1.06–1.14)	1.15* (1.09–1.21)
forg	1.12* (1.08–1.16)	1.11* (1.05–1.18)
port	1.21* (1.18–1.24)	1.22* (1.17–1.27)
comp	1.18* (1.15–1.21)	1.19* (1.14–1.23)
hum	1.12* (1.08–1.17)	1.09* (1.04–1.15)

Odds ratios correspond to the interaction terms from logistic difference-in-differences models estimated on the balanced panel (11,800 websites; 23,600 policy–snapshot observations). Standard errors are clustered at the website (`origin`) level.
* indicates statistical significance after Benjamini–Hochberg [13] false discovery rate correction at $\alpha = 0.05$, applied separately to `CH` vs. `EU` and `CH & EU` vs. `EU`.

Table 20: Logistic difference-in-differences.

The interaction terms capture differential temporal change relative to the EU baseline on the odds scale. Table 20 reports the resulting odds ratios. The qualitative pattern mirrors the primary results: interaction effects for the `CH` and `CH & EU` groups are positive and statistically significant for all obligations except `purp`. As in the primary specification, we control the expected false discovery rate using the Benjamini–Hochberg procedure at $\alpha = 0.05$, applied separately to the two families of interaction terms. The logistic specification confirms that the differential increases documented in Section 5 do not depend on the linear probability formulation.

A.11 Sample Generator Interface (SwissAnwalt)

Figure 6 displays part of the translated user interface for the most widely used (yet no longer operational) generator in our dataset, "SwissAnwalt" (retrieved using the Wayback Machine [72]). To create a policy, a user needed to tick all boxes that applied to their website, e.g., the use of a newsletter, and to provide information about the identity of the controller (not visible in the screenshot). While generating a policy was free, SwissAnwalt's business model seemed to rely on (1) using references in generated policies as advertising and (2) referring users to the network of lawyers by whom it was provided, who offered their services for a fee.

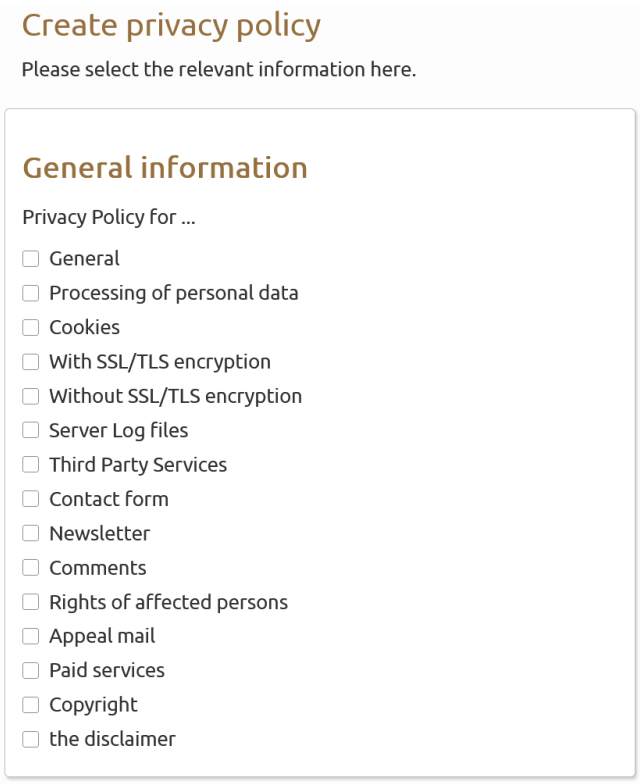


Figure 6: Translated excerpt from the latest available version of the SwissAnwalt policy generator (March 2023).

A.12 Detailed Generator Disclosure Statistics

Table 21 confirms that generator use is associated with increased mean average disclosure rates *independent* from the website popularity rank (CrUX). While generators are most prevalent among less visited websites, these smaller websites also exhibit the highest disclosure increases when using a generator. Moreover, in Section 6, we have provided the disclosure metrics for the five most widely used generators in Table 8. We provide the full analysis for all generators in Table 22.

	Top 5k rank		Top 10k–50k rank		Top 100k+ rank	
Obligation	No generator	Generator used	No generator	Generator used	No generator	Generator used
contr	86.3%	97.9%	79.7%	86.8%	76.1%	89.3%
purp	98.0%	100.0%	98.7%	99.9%	98.7%	100.0%
rect	84.9%	89.4%	80.7%	85.1%	77.4%	82.1%
forg	86.6%	89.4%	82.6%	85.4%	79.9%	82.2%
port	65.6%	78.7%	54.6%	69.8%	55.8%	67.7%
comp	66.9%	83.0%	52.6%	69.5%	52.2%	67.1%
hum	34.6%	14.9%	22.6%	13.9%	20.1%	16.3%
Average	74.7%	79.0%	67.3%	72.9%	65.7%	72.1%
Total policies	1061	47	6224	1000	5028	973

Table 21: Disclosure rates per obligation by rank group and generator use (October 2023).

Generator	*contr*	*purp*	rect	forg	port	comp	*hum*	**Average**	**Policies**	**Market Share**
SwissAnwalt (CH)	92.5%	100.0%	57.8%	58.3%	45.5%	46.1%	1.4%	57.4%	657	32.7%
Datenschutzpartner (CH)	99.1%	99.7%	99.7%	99.7%	89.3%	99.7%	0.6%	84.0%	337	16.7%
PrivacyBee (CH)	99.5%	100.0%	99.5%	99.5%	98.6%	99.5%	47.7%	92.1%	222	11.0%
eRecht24 (DE)	5.8%	100.0%	98.1%	98.1%	5.8%	5.8%	1.3%	45.0%	156	7.8%
DGD (DE)	98.3%	100.0%	98.3%	98.3%	97.5%	95.0%	99.2%	98.1%	121	6.0%
BrainBox (CH)	94.2%	100.0%	76.0%	75.2%	75.2%	73.6%	1.7%	70.8%	121	6.0%
activeMind (DE)	88.8%	100.0%	98.0%	98.0%	83.7%	75.5%	21.4%	80.8%	98	4.9%
Schwenke (DE)	89.9%	100.0%	96.2%	97.5%	89.9%	89.9%	12.7%	82.3%	79	3.9%
AdSimple (DE)	92.9%	100.0%	98.2%	100.0%	100.0%	35.7%	25.0%	78.8%	56	2.8%
WeissPartner (DE)	97.7%	100.0%	100.0%	100.0%	86.4%	86.4%	2.3%	81.8%	44	2.2%
Iubenda (IT)	100.0%	100.0%	79.1%	79.1%	79.1%	79.1%	14.0%	75.7%	43	2.1%
LegallyOK (CH)	100.0%	100.0%	100.0%	100.0%	89.7%	100.0%	15.4%	86.4%	39	1.9%
MeinDatenschutz (DE)	100.0%	100.0%	100.0%	100.0%	87.1%	77.4%	6.5%	81.6%	31	1.5%
TrustedShops (DE)	50.0%	100.0%	100.0%	100.0%	100.0%	100.0%	25.0%	82.1%	8	0.4%

Table 22: Disclosures by individual generator and obligation for policies using a single generator (October 2023).