# Automating Governing Knowledge Commons and Contextual Integrity (GKC-CI) Privacy Policy Annotations with Large Language Models


Jake Chanenson*
jchanen1@uchicago.edu
University of Chicago
Chicago, IL, USA

Madison Pickering*
madisonp@uchicago.edu
University of Chicago
Chicago, IL, USA

Noah Apthorpe
napthorpe@colgate.edu
Colgate University
Hamilton, NY, USA



## Abstract

Identifying contextual integrity (CI) and governing knowledge commons (GKC) parameters in privacy policy texts can facilitate normative privacy analysis. However, GKC-CI annotation has heretofore required manual or crowdsourced effort. This paper demonstrates that high-accuracy GKC-CI parameter annotation of privacy policies can be performed automatically using large language models. We fine-tune 50 open-source and proprietary models on 21,588 ground truth GKC-CI annotations from 16 privacy policies. Our best performing model has an accuracy of 90.65%, which is comparable to the accuracy of experts on the same task. We apply our best performing model to 456 privacy policies from a variety of online services, demonstrating the effectiveness of scaling GKC-CI annotation for privacy policy exploration and analysis. We publicly release our model training code, training and testing data, an annotation visualizer, and all annotated policies for future GKC-CI research.



## Keywords

privacy, contextual integrity, governing knowledge commons, natural language processing, large language model, text tagging


## 1 Introduction

Privacy policies are notoriously complex and lengthy documents [46]. These policies are often written in complex language or "legalese" to obfuscate the extent of data collection and discourage consumers from closely interrogating their privacy implications [1, 38, 59]. Most consumers therefore choose to ignore privacy policies when agreeing to online terms and services [62]. Even experts have difficulty interpreting some privacy policies [60]. However, privacy policies remain essential to Internet privacy broadly and to the privacy-relevant behaviors of online services.

The continued importance of privacy policies has motivated substantial research into structured methods of privacy policy analysis. Some of these methods seek to provide clearer or more easily digestible information to consumers or developers [4, 15, 63, 86], while others facilitate academic studies of the policies themselves, their relation to company behavior, or to privacy regulation [2, 3, 44, 52, 75, 89]. Both approaches often employ **annotation**—labeling relevant parts of privacy policy texts with metadata—as a primary technique.

Early successful efforts involved annotating privacy policies with a large set of metadata tags [86]. A more recent approach [67] has leveraged the theory of **contextual integrity (CI)** [48] to annotate privacy policies. CI annotation uses a small set of theoretically grounded tags to facilitate comparative and longitudinal analysis of data handling practices and policy ambiguities [67]. Research using CI as a theoretical foundation is widespread across topics in security and privacy, human-computer interaction, formal modeling, and legal analysis [7, 12, 14, 23, 68, 85]. Unlike differential privacy and other *probabilistic* privacy definitions, CI extends a *normative* notion of privacy that is a more natural fit for natural language documents like privacy policies [48].

Recent work has shown that CI is even more effective if expanded using the governing knowledge commons framework (GKC) [30, 64]. GKC provides an institutional grammar for describing strategies, norms, and rules around shared knowledge resources. The unified **GKC-CI framework** [69] (Section 3) enables straightforward identification of privacy policy ambiguities that reduce interpretability and provide excessive leeway for behavior users may consider privacy-violating. GKC-CI also enables normative analyses of contextual information transfers, the rules-in-use, and the rules-on-the-books that govern data handling practices.

All previous uses of CI parameter annotation for privacy policy analysis have involved human effort by experts or crowdworkers. Manual annotation by expert researchers produces high-quality results, but the process is tedious and slow. Crowdsourcing produces annotations more quickly, but there is a significant rate of poor-quality annotations since the annotation task is inherently nuanced [67]. Combining multiple crowdsourced annotations through a voting process can improve overall performance but further increases expense, as multiple crowdworkers must be hired to annotate overlapping sections of privacy policy text [67]. A prior study spent approximately $200 for crowdworker annotation of only 48 *excerpts* from 16 privacy policies [67].

Furthermore, no previous work on automating privacy policy annotation with machine learning [2, 3, 8, 33, 72, 74, 87, 92] has been based on CI or GKC-CI. Rather, this prior work has focused on different sets of annotation tags that are not useful to the GKC or CI scholarly communities as they do not produce results relevant to GKC or CI analysis. This motivates the development of novel automated techniques for the GKC-CI annotation task.

*Both authors contributed equally to this research.




*Proceedings on Privacy Enhancing Technologies 2025(2), 280–308*





https://doi.org/10.56553/popets-2025-0062

In this paper, we train a variety of large language models (LLMs) to perform automated GKC-CI parameter annotations of privacy policies. Doing so demonstrates the feasibility of large-scale longitudinal and cross-industry analysis of privacy policies using the GKC-CI framework. Specifically, we train and evaluate 50 LLMs from five different model families, ranging from open-source to proprietary models. We perform ablation studies with the LLMs, and briefly remark on some key lessons which we believe may be broadly applicable to other researchers working in this space. We then proceed to a nuanced discussion of the behavior of our best performing model, particularly with respect to the model's errors.

We observe that of the 50 models we benchmark, a version of `GPT-3.5 Turbo` performs the best. We find that the model boasts a robust accuracy of 90.65%, better than that of prior crowdsourcing approaches [67]. We consider an accurate annotation to be an exact string match of the human annotator's text. In contrast to prior approaches, our best performing LLM only costs $0.23-0.44 on average to annotate a privacy policy (depending on length) and is able to annotate an entire privacy policy in one minute or less on average (~5000 words a minute). We could annotate all policy excerpts from [67] for only $7 and in less than one minute.

We use our best performing model to annotate longitudinal and cross-industry policies from the Princeton-Leuven Longitudinal Corpus of Privacy Policies [1]. We evaluate these policies based on GKC-CI parameter counts, densities, and variances, demonstrating how GKC-CI annotation can be used to highlight policies of interest for further analysis. We also present a Python-based GUI that visualizes the results of GKC-CI annotation, showing privacy policy text with annotated parameters highlighted by color.

In summary, this paper presents a practical method for analyzing privacy policies that integrates machine learning, privacy law, and governance. This is the first automated method for GKC-CI parameter annotation in legal documents, which will be extremely useful to the growing GKC and CI research communities. This approach could also encourage collaboration and open up new research opportunities in these interconnected fields, promoting more in-depth exploration of privacy protection and compliance. More specifically,

- We demonstrate that accurate GKC-CI parameter annotations of privacy policies can be performed automatically by a fine-tuned large language model (LLM), substantially improving scalability and reducing expense compared to manual and crowdsourcing approaches. Further, through training data review, we find these annotations are as accurate as those obtained from humans.
- We train and publicly release the data and scripts necessary to reproduce a large language model capable of performing automatic GKC-CI annotation of privacy policies. Our automated approach has an average per-policy annotation cost between $0.23-0.44, depending on the length of the policy.[1] We note that reproducing our model incurs a one-time training cost of $195 at time of writing.
- We perform a large-scale longitudinal and cross-industry analysis of privacy policies using our best performing model. We demonstrate that the annotations can highlight policies with atypical parameter densities and distributions that may be good candidates for future in-depth evaluation. We compile all 456 annotated policies into a GitHub repository, which we make publicly available.[2]

## 2 Related Work

Substantial prior research has focused on systematic analyses of privacy policies. These analyses were often intended to improve consumer understanding of data handling processes and facilitate academic study of Internet privacy trends. This paper builds on this foundation, contributing to the broad goal of developing a library of effective, scalable, and inexpensive privacy policy analysis techniques suitable for a range of applications.

### 2.1 The Usable Privacy Project

The Usable Privacy Project [63] from Carnegie Mellon University is perhaps the most visibly successful application of annotation as a method for privacy policy interpretation and explanation. This project started in 2016 with a study by Wilson et al. [86], that recruited law students to manually annotate privacy policies with metadata tags such as "first party collection/use," "user choice/control," "data retention," and "data security." Wilson et al. also showed [88] that annotations produced by crowdworkers agreed with those of expert annotators over 80% of the time. This showed that crowdsourcing techniques could be used to identify paragraphs describing specific data handling practices in privacy policies.

In 2018, Wilson et al. used 115 expert-labeled policies to train logistic regression, support vector machine, and convolutional neural network models to automatically label sentences or segments of privacy policies with data practice categories [87]. Their best models had average F1 scores of 0.66 for policy sentences and 0.78 for policy segments. These techniques have been applied to over 7000 privacy policies from 2017, with results posted on the Usable Privacy Project website to inform consumers of the wide variety of information handling practices conducted by online services.[3]

This line of research, while independently impressive, is not relevant for the community of scholars using contextual integrity or governing knowledge commons to interpret privacy policies and other legal documents. The sets of annotation tags used for the Usable Privacy Project are not analogous to the 5-parameter information flow description fundamental to CI nor the institutional grammar fundamental to GKC. This means that a separate thread of research, described below, has been needed to develop privacy policy annotation approaches useful for GKC and CI analyses.

Unlike the work by Wilson et al. [86–88] and the Usable Privacy Project, our work is based on Shvartzshnaider et al.'s GKC-CI framework [69], which provides a theoretically grounded basis for identifying ambiguity and potential privacy-violating behavior. Our work is less focused on helping consumers understand privacy policies than [86–88] and is more focused on automatable, longitudinal, and cross-industry analysis of privacy policies. Our work is also

[1] This cost estimation range is the average cost of annotating privacy policies from both the 164 least popular and the 164 most popular websites, per their Tranco [53] ranking, in the Princeton-Leuven Longitudinal Corpus of Privacy Policies [1].

[2] https://github.com/JakeC007/Automated_GKC-CI_Privacy_Policy_Annotations

[3] https://explore.usableprivacy.org/

novel in its use of large language models (the current state-of-the-art in natural language processing) and achieves considerably better performance than the machine learning annotation in [87].

## 2.2 Manual CI Annotation

In 2019, Shvartzshnaider et al. [67] used the theory of contextual integrity (CI) [48] to inform a new approach to privacy policy annotation. This approach seeks to identify the five information flow parameters defined by CI (Section 3) in privacy policy text. CI parameter annotation enables the identification of ambiguities in information transfer descriptions. In 2022, Shvartzshnaider et al. [69] combined contextual integrity with governing knowledge commons (GKC) [30, 64] to create a combined GKC-CI framework. GKC-CI extends the potential scope of CI annotation to eight total parameters, four from CI and four from the GKC institutional grammar (Section 3).

While Shvartzshnaider et al. [67] successfully motivated CI parameter annotation for privacy policy analysis, questions of scalability remained. As with most annotation tasks, manual annotation by experts is highly accurate but tedious and slow. Shvartzshnaider et al. demonstrated that crowdsourcing could partially solve this problem but remains expensive, as high error rates necessitated the combination of multiple overlapping crowdsourced annotations per policy segment to increase precision. The resulting crowdsourced annotations still had a relatively high rate of false negative errors, *i.e.,* parameters missed by the majority of crowdworkers.

**The scalability issues posed by crowdsourced annotation clearly motivate this study**, which seeks to automate CI parameter annotation through the use of large language models (LLMs). Our work is also novel in its use of the expanded eight-parameter GKC-CI labels as annotation tags (Section 3) rather than the five-parameter CI tags used in Shvartzshnaider et al.'s original paper.

We apply our automated method to annotate a large corpus of privacy policies, including up to 20 years of longitudinal policies from 8 major technology companies (128 policies) and 328 contemporary policies from across the technology industry. This is orders of magnitude more privacy policies than have been manually annotated for CI research in previous work [67].

## 2.3 Privacy Policy Analysis With Machine Learning

Several other studies have also applied machine learning to privacy policies, although none have been based on nor produced output useful for GKC-CI research. In 2018, Harkous et al. [33] trained a hierarchy of convolutional neural networks to build a Question-Answering system that supports free-form querying of privacy policy content. Other ML-based approaches essentially parse privacy policies for information of interest, such as by using a logistic regression model to identify opt-out statements in privacy policy text [8]. PoliCheck [3], an expansion of PolicyLint [2], is capable of differentiating between first-party and third-party entities in flow-to-policy consistency analysis. Zimmeck et al. [92] and Story et al. [72] used support vector machines to identify non-compliance between Android application code and the applications' privacy policies. Their approach could be used to highlight these statements for consumers to make opt-out decisions without needing to read the entire policy themselves. Our application of machine learning to GKC-CI privacy policy annotation is similarly tightly focused, but on a task that does not overlap these earlier works.

More recently, some research groups have utilized LLMs in the legal space broadly or for privacy policy analysis specifically. LLM benchmarks in the legal space include that of Dai et al. [22] and Fei et al. [29]. Ravichander et al. [58] trained a BERT-based large language model to answer questions about privacy policies in a Q&A format (not annotation) using a corpus of 1750 questions and 3500 expert answers. Other work that adapts LLMs for legal question-answering include that of Wan et al. [82] and Yue et al. [90].

Tang et al. [74] used various LLMs to annotate privacy policies via prompting. However, the annotations are very simple (e.g., "1st Party Collection") and could likely be found using regexes. In contrast, our annotation approach uses a nuanced and theoretically grounded framework (GKC-CI) for normative privacy analysis. We believe LLMs are particularly suited for *nuanced* annotations as opposed to classical neural network methods. This is because of LLMs' increased representational capacity as well as their training on Internet webscrapes, which exposes LLMs to cultural norms [24, 56, 80]. However, we cannot provide a comparison to other researcher's classical approaches for automated GKC-CI annotation because there are none developed in this space. To address this, we train an RNN on the GKC-CI annotation task as a baseline for our work.

# 3 GKC-CI Theory

The theory of contextual integrity (CI) [48] defines privacy as the adherence of information transfers, or "flows," to sociocultural norms in specific contexts. For example, an information flow that might be appropriate between a patient and a doctor in a medical context (e.g., about a sensitive diagnosis) might not be appropriate between that doctor and their acquaintance in a recreational context.

CI further defines information flows as consisting of five essential parameters: 1) the *sender* of the information, 2) the *recipient* of the information, 3) the *subject* of the information, 4) the information content or *attribute*, and 5) the *transmission principle* that describes how or why the information flow occurs. The CI parameter annotation task entails identifying and labeling these five parameters in descriptions of information flows. For example, the CI annotation of: "We also collect contact information that you provide if you upload, sync or import this information from a device," would label "we" as a recipient, "contact information" as an attribute, "you" as the sender, and "if you upload, sync or import this information from a device" as a transmission principle (example from [67]).

The combined GKC-CI framework [69] further extends the CI framework, enabling the evaluation of strategies, norms, and rules drawn from theories of information governance. This allows cross-disciplinary research between the GKC and CI communities and allows a broader set of research questions to be addressed than either framework alone [69]. Relevant to annotation, the combined framework divides the *transmission principle* into four categories drawn from the GKC institutional grammar: 1) *aims* and/or goals for specific actions, 2) *conditions* indicating when, where, or how aims apply, 3) *modality* operators implying pressure (deontics) or hedging, and 4) *consequences*, including sanctions for noncompliance, penalties in absence of consent, and benefits for proceeding.

| Privacy Policy Sentence | Parameter | Annotated Text | Parameter | Annotated Text |
|---|---|---|---|---|
| *If you consent,* we *may* share information about you with companies that aggregate it *to provide analytics and measurement reports to our partners.* | Aim | to provide...our partners | Modality | may |
| | Attribute | information | Recipient | companies that aggregate it |
| | Condition | If you consent | Sender | we |
| | Consequence | N/A | Subject | you |

**Table 1: Example GKC-CI annotation. Italics and underlining added for emphasis.**

The GKC-CI parameter annotation task is identical to the CI annotation task except that it requires identifying the eight GKC-CI parameters instead of the five CI parameters. GKC-CI annotations thus provide more nuance than CI annotations at the expense of increased annotation difficulty. An example of what different GKC-CI parameters are present in a sample sentence is shown in Table 1. We encourage readers interested in additional information about the GKC-CI framework to refer to the original paper [69], which explains the theoretical foundations in detail and provides a worked demonstration of GKC-CI analysis in the Internet of things context.

## 4 Methods

We next describe how we trained LLMs to accurately perform GKC-CI parameter annotation for privacy policies. In doing so, we make the following contributions.

First, we tested how various architectures and sizes of different LLMs effect performance through ablation experiments. After comparing a broad selection of models to find the most suitable candidate, we further tuned the hyperparameters of our best performing model. Details of the model training process are provided in Section 4, and performance results are provided in Section 5. Because there is no previous work demonstrating how NLP models may be applied for GKC-CI annotation, we believe that these results will be particularly helpful to those considering normative privacy policy analysis in this space.

Second, we qualitatively examined the errors made by our best performing model to gain insights into the model's behavior, particularly as it relates to practical deployment. We believe that our approach to error analysis may provide a useful foundation for how others may analyze model behavior in this space. The details of our error analysis are reported in Section 5.4.

Finally, we used our best performing model to annotate a large set of privacy policies [1], which we report in Section 7, providing examples of the types of longitudinal and cross-industry analyses that can be performed via automated GKC-CI annotation.

### 4.1 Training and Testing Data

Our ground-truth labels were obtained by manually annotating GKC-CI parameters in 16 privacy policies from popular online services and e-learning websites, the exact breakdown of which is shown in Table 3 of Appendix A. We downloaded these privacy policies in HTML format and converted them to plain text for annotation. We used a customized version of the Brat Rapid Annotation Tool [71] to manually label all GKC-CI parameters in the policies. In order to achieve consistent annotations across all annotators, we used a fixed set of guidelines defining each of the GKC-CI parameters (Appendix B). These guidelines were taken from [67] for CI parameters and [69] for GKC-CI parameters to ensure continuity with prior work. Our ground-truth annotations included 6781 GKC-CI parameters across all 16 policies (Table 3). This ground-truth annotation process took two research assistants one semester to perform, including time spent learning the task.

In the process of annotating, we encountered several of the challenges discussed in [67], including implicit parameters, ambiguous parameters, and policies not written with the CI framework in mind. We addressed these issues consistently with [67]. In general, the annotators made best judgment calls when faced with ambiguous parameters or difficult logic, consulting with the authors to ensure consistency. Importantly, we did not expect these manual annotations to be perfect. Rather, we treated them as best-effort annotations by researchers familiar with the task.

### 4.2 Formatting Examples

We lightly formatted each sentence of our ground-truth annotated privacy policies as the basis of our training and testing examples. Each formatted example consisted of the following parts: (1) a prefix to orient the LLM to the task, (2) a sentence from a privacy policy, (3) the GKC-CI parameter of interest, and (4) text delimiters. We chose the extremely minimal prefix "Annotate:" to minimize the effects of prompt choice while still leveraging the training benefits of using a prompt [65, 83]. We included the text delimiters because modern LLMs decide what text to generate next based on *all* the text in their context window. As such, they cannot by default determine what text has been provided via the prompt and what the LLM has generated. We chose our text delimiters based on recommendations in OpenAI's documentation, namely "–>" and "x-x-x" respectively. Two fully formatted examples are shown below:

(1) Annotate: ["We also collect contact information that you provide"] Recipient–> Recipient: ["We"]x-x-x
(2) Annotate: ["We also collect contact information that you provide"] Aim–> Aim: N/Ax-x-x

In formatting these examples, we wanted to ensure that the models learned to find the relevant text to be annotated. This necessitated teaching the models to filter through *irrelevant* text. As such, we used *positive examples* (text where a parameter is present, such as example (1) above) and *negative examples* (text where a given parameter is not present, such as example (2) above). The inclusion of negative examples was necessary to ensure that the models are usable in a real-world environment; not every sentence of a privacy policy will include a GKC-CI parameter. By including negative examples, our models learned to only output a parameter if one is actually present in an input sentence. We created the negative examples by taking positive examples and changing the parameter of interest to one which does not occur in the sentence. We did this for each positive example and all possible parameters not already in that example. For instance, the negative example

(2) above was created by switching the parameter of interest in the positive example (1) above from "recipient" (which should be labeled in the example text) to "aim" (which should not).

We used sentences as the atomic unit for model input because sentence divisions are natural delimiters and previous work has shown reasonable accuracy with sentence-based annotation [87]. We observed that other natural units of separation, such as paragraphs, tended to be particularly long due to the legal nature of privacy policies. This length ultimately resulted in problems where the paragraphs were too long to fit into the context window of some of our LLMs.[4] To ensure that the training data was consistent between LLMs, and because of the prior work above, we used sentences as the basis of our annotations. However, we also performed a supplemental analysis with GPT-3.5 to verify that we were not critically disadvantaging our models by only feeding in sentences (Appendix C).

*4.2.1 OpenAI Models.* Some OpenAI chatbot models take an additional prompt, a *system message*, as input. GPT-3.5 Turbo is one such model. For these models, we replaced the default system message from *"You are a helpful assistant."* to *"You are an assistant that understands Helen Nissenbaum's theory of Contextual integrity (CI) and the governance of knowledge commons framework (GKC). This framework is abbreviated as GKC-CI. You reply with brief, to-the-point answers with no elaboration."* We did so to help orient the model to our task.

Finally, we performed an experiment leveraging the fact that GPT-3.5 Turbo is designed to respond to *conversational* inputs, as it is a chatbot. Specifically, we changed the prefix from *"Annotate:"* to *"For the following excerpt, provide the GKC-CI annotation of '<parameter>':"*. We call the model produced under this intervention `GPT-3.5 Turbo, Prompt Engineered`. Note that this is distinct from our baseline models that we prompted *without fine-tuning*.

## 4.3 Model Selection

*4.3.1 Baseline Models.* We considered two baselines for comparison against our fine-tuned LLMs: 1) a recurrent neural network (RNN) representing classical NLP approaches and 2) prompted *non-fine-tuned* LLMs.

For the RNN, we defined BOS and EOS tokens, which is standard practice for these models [61]. We then perform standard training for the RNN but keep the rest of the training parameters consistent with the LLMs as described in Section 4.4.

For the prompted non-fine-tuned LLMs, we used GPT-4, GPT-4 Turbo, and GPT-3.5 Turbo. We used a fixed prompt format where we clearly delineated instructions from other text by placing instructions within brackets, thereby providing a structured framework that helped the model distinguish between the task directives and supplementary information. We also employed n-shot learning, a common prompting technique that has been shown to improve performance as n, the number of examples, increases [16, 31, 91]. We used n={1, 3, 5} where the examples were chosen randomly from the training set.

[4]Further, to our knowledge, resolving a finite-length context window with large quantities of text is an open problem in the NLP space, although there has been substantial work in the area, such as [54] and [11].

*4.3.2 Fine-Tuned Models.* We considered five model families of diverse size and architecture: Flan-T5, GPT-2, Llama2, GPT-3, GPT-3.5 Turbo, and GPT-4 [16, 19, 57, 78]. Their properties are summarized in Table 2. In selecting models from these families, we wanted to choose from a wide range of high-performing or particularly usable LLMs. We ultimately omitted GPT-4 from our analysis because at the time of writing, the fine-tuning API was experimental. We also omitted Llama2's chat version from our analysis because, at the time of writing, it did not appear that Meta intended it to be fine-tuned further based on its documentation [28]. Of the models we selected, approximately half are open-source, while the GPT-3 and GPT-3.5 models are proprietary. While the exact sizes of the GPT models are not publicly released, they are likely the among the largest of the models we tested.

We now give a quick summary of how the various architectural features of the models we trained may impact performance on the annotation task. First, we varied model size. Specifically, we considered models both in their "base" or default size[5] as well as a larger size, if permissible by our hardware.[6] We considered model size because the number of parameters in a model plays a large role in its performance [19, 34, 39, 77]. However, smaller models may approach larger models' performance if they are given more (pre)training data. We consequently included the Llama2 model to serve as an open-source proxy for larger models' (e.g., GPT-3) performance [77, 78].

Most newer LLMs employ a "decoder-only" architecture, a change from the original design of the Transformer [80]. Such models are referred to as encoder-decoder models, while models lacking an encoder block are decoder-only. Decoder-only models can only "look" at the preceding tokens to determine what should be generated next and tend to excel at creative or free-form generation [35]. In contrast, encoder-decoder models look at the entire input sequence to determine what should be generated. Encoder-decoder models, like Flan-T5 or BERT, tend to perform well on tasks where the output is highly scoped by the input [19, 24, 35]. We included both decoder-only and encoder-decoder models in our analysis because it was unclear what advantage, if any, an encoder-decoder model might have when annotating privacy policies.

There exist a number of training paradigms which result in a model becoming more **aligned** to human intent. We specifically mean that a model is aligned if it produces outputs which are consistent with its human operator's desires (assuming the model is capable of producing those outputs) [18, 49]. Because alignment broadly reflects a model's ability to output text consistent with input tasks, we hypothesized that models which have undergone specific alignment training might perform better. We considered two alignment mechanisms: Instruction-Finetuning and Reinforcement Learning with Human Feedback (RLHF) [49, 84]. We included models in our analysis which are either confirmed or likely to have been trained according to these techniques.

Finally, we noted that some models have been released as chat models. This is relevant because 1) such models can be prompted in a different format from other non-chat models (Section 4.2.1), and 2) these models are generally newer. While we are loath to conflate

[5]When loading from the HuggingFace model hub

[6]We employed an NVIDIA A100 for training, which we believe to be a reasonable baseline for the amount of compute other researchers will have available.

| Model Family | Open-Source | Size(s) Considered | Architecture | Instruction-Finetuned | RLHF | Chat Variant |
|---|---|---|---|---|---|---|
| GPT-2 | Yes | Base (124M), XL (1.5B) | Decoder-Only | No | No | No |
| Flan-T5 | Yes | Base (248M), Large (783M) | Encoder-Decoder | Yes | No | No |
| Llama2 | Yes | 7B | Decoder-Only | No | No | No |
| GPT-3 | No | "Davinci" | Decoder-Only | Unconfirmed | Unconfirmed | No |
| GPT-3.5 Turbo | No | Unknown | Decoder-Only | Unconfirmed | Unconfirmed | Yes |

**Table 2: A summary of the models we trained and their model families.**

model age with performance, it is at least in the case of GPT and Llama that newer releases tend to eclipse older releases [16, 57, 77, 78]. We therefore included the fine-tunable chat model GPT-3.5 Turbo in our analysis.

## 4.4 Model Training

We processed all 16 manually-annotated privacy policies into ground truth examples for model training and testing as described in Section 4.2. We randomly reserved 70% of the manual annotations to constitute our training data (21,588 examples), while the other 30% (9252 examples) were testing data.

We used low-rank adaptation (LoRA), a type of parameter efficient fine-tuning (PEFT), as our training method to ensure that the open-source models were trained in a way consistent with the proprietary models [37]. PEFT describes a number of methods for training some, but not all, of a model's parameters (i.e., being "parameter efficient" when training). LoRA is a PEFT method that involves freezing a model's weights and inserting trainable rank decomposition matrices into the model's architecture. These low-rank matrices are then "trained" during fine-tuning, while the base model remains frozen. OpenAI's business model suggests that LoRA is being employed in the place of traditional fine-tuning, as low rank matrices are very cheap to store and entire models are typically large and expensive.[7]

Thus, to ensure that all our model comparisons are fair, we trained all LLMs using LoRA. We kept the training parameters constant between all open-source models. We made an exception for the learning rate of Flan-T5, as the model's documentation recommends a slightly higher learning rate than the other models. This higher learning rate could have caused the model to converge faster than others.

During training, we performed a number of ablation experiments, changing the following in a full-factorial manner for the open-source models:

(1) **Training Epochs**. We varied the number of training epochs for the model. Training a model for a single epoch means that the model sees every example in the training set once. Training for $n$ epochs means that the model sees every example in the training set $n$ times. Varying the number of training epochs thus influences the quantity of training tokens to which the model is exposed, as well as the number of repeated examples to which the model is exposed.
(2) **Example formatting**. We varied the examples' formatting such that all training examples were prepended and appended with the model's BOS, EOS tokens or not depending on the ablation condition. BOS (Beginning of Sentence) and EOS (End of Sentence) tokens are typically defined and used during LLM pre-training to semantically capture the start and end of a training document. We hypothesized that the inclusion or exclusion of these tokens during fine-tuning may have an effect on the model's performance.

OpenAI's proprietary models do not offer the same number of training options, so while were were able to experiment with different numbers of training epochs for OpenAI's proprietary models, we were unable to perform the example formatting experiment described above with the proprietary models. We ultimately benchmarked 50 models distributed between open-source and OpenAI models.

# 5 Model Performance

## 5.1 Metrics

We begin our discussion of performance by clearly defining our metrics. Specifically, our main metric is *accuracy*. We consider a model's output to be correct or *accurate* if it is an exact string match of the ground truth. We do not directly consider precision and recall because they are less intuitive for this task. We envision that downstream users are likely to be particularly interested in the performance per GKC-CI parameter. As such, we focus on per-parameter accuracy.

Since we want to capture nuanced model errors, we categorize each model output into exactly one of four possible results: *perfect match*, *superset match*, *match error*, or *identification error*. These categories are defined as follows:

(1) **Perfect Match** indicates that the model's annotation is an *exact string match* with that of the human annotator. *Only a Perfect Match represents a correct annotation.*
(2) **Superset Match** indicates that the model's annotation contains all the words of the human annotator. However, the model may have highlighted additional information or may have included information which does not appear as a contiguous sequence of text in the policy.
(3) **Match Error** indicates that the model agreed that a certain parameter was present but did not identify the "correct" annotation. This can include completions that are flat-out incorrect, completions that don't identify the correct number

[7]We make this claim because OpenAI allows for "fine-tuning" through their API. Traditional fine-tuning would require making full copies of the model (*e.g.*, GPT-3.5) for each user. Doing so would result in terabytes of space being allocated per user due to the size of OpenAI's models. Additionally, it has been observed that previously fine-tuned OpenAI models may change in performance without warning, as OpenAI routinely updates their models. This behavior would not be observed if traditional fine-tuning were occurring because each user would have their own discrete copy of the model.

of instances of a parameter in the input, and completions that have identified a proper subset ($\subset$) of the correct words.

(4) **Identification Error** occurs when the model, despite being prompted with a specific parameter (*e.g.,* "Aim"), failed to include that parameter in its completion.

Because any model output is categorized as belonging to one of the above categories, accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Correct Outputs}}{\text{All Outputs}} = \frac{PM}{PM + SM + ME + IE}$$

where $PM$ is the number of perfect match annotations, $SM$ is the number of superset matches, $ME$ is the number of match errors, and $IE$ is the number of identification errors.

## 5.2 Baseline Results

Our baseline RNN model had a test set accuracy of only 6% (Figure 1). While the RNN handled negative examples well, it struggled with positive examples. Specifically, the RNN failed to accurately annotate *any* positive examples, while all LLMs other than GPT-2 were able to successfully annotate at least one positive example.

The prompted *non-fine-tuned* LLMs also performed poorly, with the best such model having a test set accuracy of <20% (Appendix D). While more effective prompting strategies are likely to exist, we believe that this finding provides evidence that LLMs should not be applied off-the-shelf for GKC-CI annotation as of now.

Comparing these results to the performance of the fine-tuned LLMs below, it is evident that fine-tuned LLMs substantially outperform both classical NLP RNNs and non-fined-tuned LLMs. This justifies the time and expense needed for fine-tuning and further indicates the non-triviality of the GKC-CI annotation task.

## 5.3 Fine-Tuned Results

We benchmarked each of our 50 models on each of the 9252 sentences in our test set. The results are shown in Figure 1. OpenAI's proprietary models performed significantly better than any of the open-source models we tested. No open-source model that we tested had performance high enough to be considered even a poor substitute. Examination of the open-source models' outputs indicates a lack of substantial training convergence within 10 epochs for all models. Graphs of the open source models' performance at 1, 5, and 10 epochs are included in Appendix F. Each of the open-source models took <12 hours to train, but the training time is extremely sensitive to specific hardware configurations. We hypothesize that given enough compute, modern open-source models may eventually reach the performance of OpenAI's proprietary models. We briefly summarize some notable lessons from working with the open-source models below. We then discuss the performance of the OpenAI models in greater detail.

*5.3.1 Open-Source Takeaways.* First, we note that within a model family, size does play a significant role in a model's ability to perform well on our annotation task. GPT-2 and Flan-T5 both failed utterly at their smaller model sizes, while their XL and Large variants performed $\approx$3% to 15% better. We additionally note that absolute model size in terms of parameters across model families does not appear to be a consistent indicator of performance. Llama2's 7B variety performed similarly to GPT-2's XL variety, despite being over

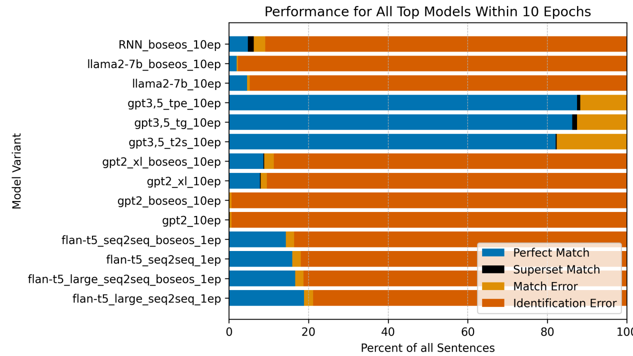


**Figure 1: Test set performance of the top-performing models variants, including the RNN, with $\leq$ 10 epoch of training. `GPT3,5_TPE` refers to the prompt-engineered version of `GPT-3.5 Turbo`, `GPT3,5_TG` refers the generic GPT-3.5 Turbo model, and GPT3,5_t2s refers to the joint performance of the GPT-3.5 Turbo, 2-Step models. Expanded model names in Appendix E.**

four times the size. We mention this to caution resource-constrained researchers: **larger may not always mean better when comparing across model families.** Additionally, those models which were aligned (Flan-T5 and all GPT-3.5 models) performed the best. We **recommend compute-constrained researchers prioritize smaller, aligned models.**

Second, we observe that models appear to be vulnerable to the inclusion or exclusion of BOS, EOS tokens. Namely, we observe small performance differences between *all* open-source models, but most strikingly in the Llama2 model family. We believe this is significant because it suggests that opaque defaults have observable effects on performance. Further, those effects do not appear to be consistent across model families. Specifically, as of the time of writing, HuggingFace's tokenizers *all* have BOS, EOS tokens internally defined, but each model has a different default behavior when it comes to including or excluding BOS, EOS tokens. **We urge other researchers to be particularly careful of library and model defaults as they could be a potential confound in model performance.**

*5.3.2 OpenAI Models.* Appendix G summarizes model performance for three variants of GPT-3.5 Turbo on *positive* examples, which are those examples that contain a GKC-CI parameter. We observe that these models vary substantially in their per-parameter performance.

`GPT3.5-Turbo, two step 10 epochs` refers to a two model system where the first model determines if the text contains a GKC-CI parameter and the second model identifies the parameter. For this model, an example is a perfect match if and only if the first model and the second model classify the prompt correctly. This model performed the worst of the three GPT-3.5 Turbo variants with 30% correct positive examples and 97% correct negative examples. While this model demonstrated success identifying negative examples, its limited capacity to accurately identify positive examples significantly undermined its overall performance.

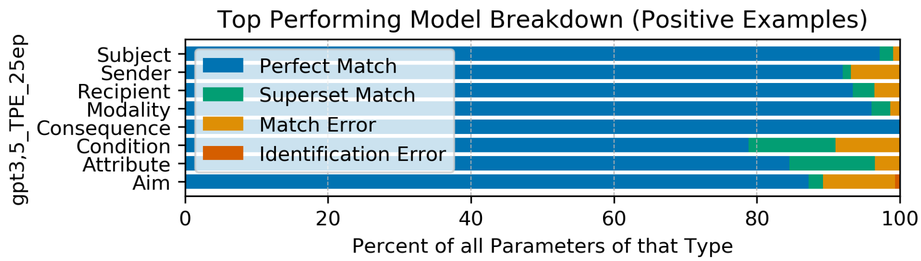


**Figure 2: Performance per GKC-CI parameter for our best performing model, `GPT 3.5TPE_25ep`.**

`GPT3.5-Turbo, Generic 10 epochs` (`GPT 3.5TG_10ep`) and `GPT3.5-Turbo, Prompt Engineered 10 epochs` (`GPT 3.5TPE_10ep`) performed similarly well to each other. `GPT 3.5TG_10ep` had 73% correct positive examples and 88% correct negative examples. `GPT 3.5TPE_10ep` had a slightly better 75% correct positive examples and 89% correct negative examples.

We speculate that the reason why `GPT 3.5TPE_10ep` performed better than `GPT 3.5TG_10ep` is due to the added context from prompt engineering. Specifically, prompt engineering may provide enough context for `GPT 3.5TPE_10ep` such that the model is able to relate the annotation task to information on which it has been trained. This phenomenon has been more broadly observed in the NLP space [65, 83].

We next investigated how the number of training epochs effected the performance of `GPT 3.5TPE` (Figure 3). We observed that `GPT 3.5TPE` started to have diminishing returns after 25 epochs worth of training. This behavior is typical for any model trained using gradient descent with an optimizer since both the steps grow smaller as the model approaches a minima and the learning rate decays over time. We hypothesize that the model had likely overfit after 50 epochs. We further investigated the effect of training data quantity on `GPT 3.5TPE` (Figure 4). The model started to have diminishing returns in terms of performance per data quantity when using 75% of the training data, and 50% of the training data appeared to be the necessary minimum for the model to begin to converge. However, we observed no detriment to using all available training data.

To summarize, we observed that `GPT 3.5TPE` benefited from the full training dataset and that the model was at risk of overfitting after 25 epochs. **We consequently chose `GPT 3.5TPE` as our best performing model and trained it for 25 epochs using the full training data**, which took 3 hours. **We call this top model `GPT 3.5TPE_25ep`**, and its performance across GKC-CI parameters in the test set is shown in Figure 2. To better understand the strengths and weaknesses of this model, we performed a qualitative analysis of the model's errors (Section 5.4).

Finally, to ensure that this model's performance isn't critically hindered by being trained exclusively on sentences, we fine-tuned a version of `GPT 3.5TPE` using paragraphs instead of sentences (details in Appendix M). The goal was to determine whether providing more context in the input would improve performance. Despite subjecting this model to the same training epoch ablation experiments as our top-performing model, it only achieved a maximum accuracy of 67.75%, substantially lower than the 90.65% maximum accuracy achieved with sentence-based training. For a detailed

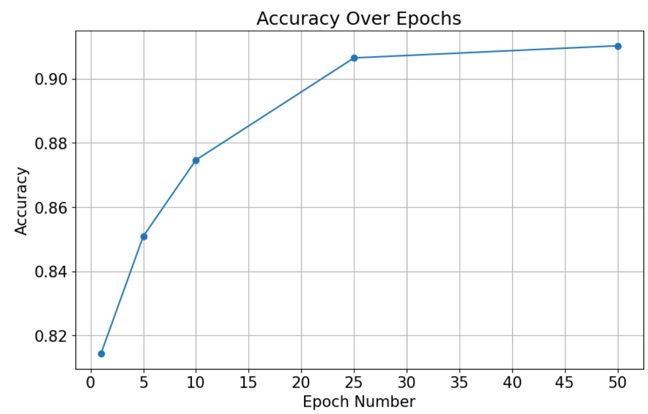


**Figure 3: `GPT 3.5TPE`'s performance on the test set at 1, 5, 10, 25, and 50 epochs. Only Perfect Matches were considered to be "correct."**

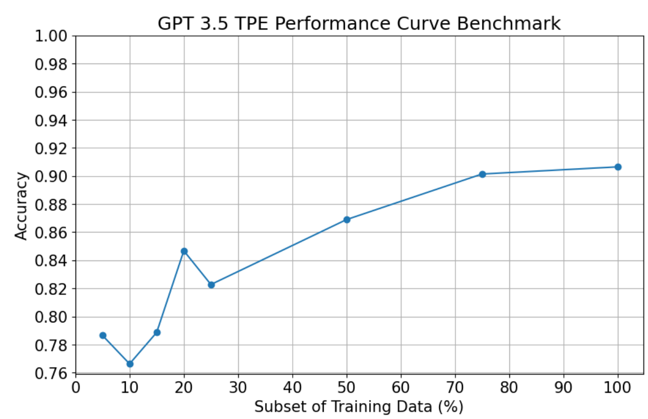


**Figure 4: `GPT 3.5TPE_25ep`'s accuracy on the test set as training data increases. Only Perfect Matches were considered to be "correct."**

comparison of the performance between these sentence-based and paragraph-based models, please refer to Appendix C.

## 5.4 Qualitative Error Analysis

In order to better understand the errors made by `GPT 3.5TPE_25ep`, we performed qualitative coding on the 37 *match errors* for positive examples produced by the model, i.e., match errors where neither the ground truth text nor the model completion was "N/A". This served two purposes.

First, we identified cases where model outputs were mistakenly annotated as errors, specifically where the model annotation was *semantically equivalent*, albeit *syntactically different* from the ground truth. Second, we conducted a detailed analysis of the model's performance, examining the factors contributing to both its strengths and weaknesses. This provided more confidence in overall model performance.

Identifying trends in these errors offers valuable insights into the model's behavior. Although these insights may not explain *why* an error occurs (LLM models have notoriously poor explainability [9, 26]), they help users become aware of the model's limitations.

*5.4.1 Qualitative Coding.* To ensure the reliability and consistency of the coding process, two expert coders initially met to collaboratively develop a comprehensive codebook consisting of ten codes: three parent codes and seven child codes. The full codebook can be seen in Appendix H.

After joint codebook creation, each coder independently coded all match errors produced by `GPT 3.5TPE_25ep`. Once the coding was complete, we computed inter-coder reliability and found a high level of agreement between the two coders with a Cohen's kappa score of 0.87 [20]. The results of the qualitative coding are visualized in Figure 5 and detailed further below. Note that in the following text, parent codes from the codebook are italicized, while child codes are enclosed in quotation marks

*5.4.2 Semantic Equivalence.* The child code "Semantic Equivalence," the sole child code of the parent code *Semantic Equivalence*, was the most prevalent code in our error analysis—accounting for 19/37 (51%) errors. The following two examples demonstrate this type of error. The text in quotation is the expert annotation, while the underlined sections are the model's annotation:

(1) Aim: "to <u>help us operate or administer the Services</u>"
(2) Recipient: "These <u>Services</u>"

Note that the model's response only differed by an article or an adjective, and both are equivalently correct annotations.

*5.4.3 Incorrect Expert Annotations.* The parent code *Expert Labels Is Wrong* had nine examples (24%). Occasional expert mis-annotations are expected for a task of this complexity. We are encouraged that there were relatively few examples under this parent code, supporting the quality of our ground truth. Importantly, for the examples in the "Expansive Ground Truth" (4/9, 44%) and "Partial Ground Truth" (3/9, 33%) child codes, *the model performed the task more correctly than the expert annotator*—either by omitting superfluous words included in the expert annotation or including necessary ones the expert annotator missed.

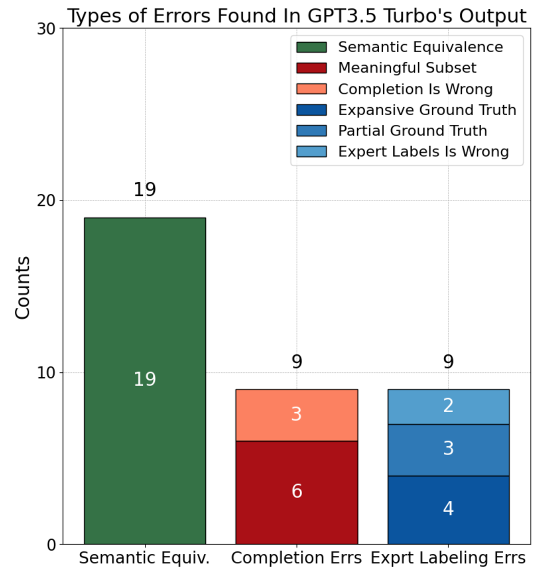


**Figure 5: Breakdown by parent code of the various types of errors found from our qualitative analysis.**

Consider the following example "Expansive Ground Truth" annotation. The text in quotation is the expert annotation, while the crossed out section is what the model correctly excluded:

(1) Consequence: "~~You can set your browser to not accept cookies, but~~ this may limit your ability to use the Services."

Conversely, the following example "Partial Ground Truth" annotations show the expert annotation in quotations, while the underlined words are what the model correctly choose to additionally include in its response:

(1) Recipient: "trusted companies" <u>that work with, or on behalf of, Crowdmark to process information</u>
(2) Condition: "to comply with its general obligations under the GDPR," <u>in particular to process the personal data it collects in accordance with Articles 5 and 6, and to comply with Articles 13, 14, 24, 30 and 32, and to comply with any actionable rights of the data subject</u>

Combined, "Expansive Ground Truth" and "Partial Ground Truth" codes represent only 9 of the 37 errors (24%); however, we emphasize this finding as particularly exciting because they demonstrate that the model can identify precise annotations for the requested parameter. In other words, the model's ability has surpassed that of our trained human annotators in these situations.

*5.4.4 True Model Errors.* The parent code *Completion Errors* accounted for 9 out of 37 coded examples (24%). Notably, a significant majority of these errors (6 out of 9, 67%) fall under the "Meaningful Subset" child code. A "Meaningful Subset" annotation included a segment of the correct response, but missed words that altered the meaning of the annotation. In the following example, the text in quotation is the expert annotation while the underlined sections are the model's annotation:

(1) Aim: "solely for the purposes of providing the relevant services to Kaltura"
(2) Attribute: "personal data, any communications or material of any kind that you e-mail, post, or transmit through the Site, such as questions, comments, suggestions, and other data"

Many of these completions do encompass enough of the correct answer for someone well-versed in GKC-CI to grasp the intended annotation. However, we consider these incomplete responses to be incorrect even though the model's answer is a meaningful subset of the correct response.

Finally, the parent code *Completion Is Wrong* comprises the final 3 out of 9 codes (33%). All of these responses annotated text from within the sentence with no relation to the actual GKC-CI parameter. For example the model incorrectly annotates the underlined part of the following sentence as an *attribute*:

(1) Attribute: "from the institution including the user's identifier and organizational affiliation"

Our qualitative analysis offers a comprehensive view of the model's performance, detailing both its strengths and limitations. We find that appreciably more than half of the purported match errors are the codes "Semantic Equivalence," "Expansive Ground Truth," and "Partial Ground Truth." These combined child codes constitute 27 examples, resulting in a 0.29 percent *increase* in the number of correct annotations overall. This implies that our benchmarking metric of 90.65% model accuracy serves as a conservative estimate of model performance.

### 5.5 Training Data Review

Motivated by the nine expert annotation errors we uncovered in Section 5.4.3, we had two different CI experts review a random sample of 5% of our training data. These experts agreed with the training data on 90.57% of the sample, with the most common disagreement being the identification of occasional *attribute* parameters that were missed in the original annotations. The original annotators spent several months annotating thousands of parameters, so it is unsurprising that a few were missed. Furthermore, governing knowledge commons and contextual integrity are the topic of active research and subject to different interpretive nuances by different members of the community, which is reflected in different annotations. This training data review validates our results, as our best performing model's 90.65% accuracy is close to the inter-expert annotation agreement, suggesting that the model approaches human performance on the task.

## 6 Limitations

Despite the high accuracy of our best performing model, we point out a few limitations. First, website privacy policies are not written with the GKC-CI framework in mind, so a perfect mapping from text to GKC-CI parameters may not be possible. We do not know the maximum performance that can be achieved by GKC-CI annotation, and it likely varies depending on the specific documents annotated. While we have created and used the largest training set for this task to date, and our analysis indicates that further increasing training set size would not substantially improve performance (Figure 4), it is possible that training data from a different set of policy documents might affect overall performance.

Second, the LLMs we use do not have information about the input text outside of the context window. This is unlike a human lawyer, who could review the entire document while identifying information flow descriptions. Fortunately, the information needed to identify GKC-CI parameters typically appears in the text immediately surrounding the parameter itself. The decreased accuracy we observed when using a longer context window (Section 5.3.2, Appendix C) supports our use of one-sentence contexts.

Third, we do not perform coreference resolution, i.e., identifying when different sections of text refer to the same real world entity. For example, a sentence may refer to the website user multiple times, with each instance receiving a GKC-CI parameter annotation (e.g., "sender"). Although our model would accurately annotate each mention of the sender, it would not automatically recognize that these annotations refer to the same legal entity. If that were necessary for a downstream application, another model would be required. Since coreference resolution is an active research problem in its own right, and is independent from GKC-CI parameter annotation, we leave this for future development.

## 7 Example Applications

We applied our `GPT 3.5TPE_25ep` model to 456 privacy policies from the Princeton-Leuven Longitudinal Corpus of Privacy Policies [1] to demonstrate the type of analyses enabled by GKC-CI annotation at scale. This dataset contains over 1 million privacy policies from over 100,000 companies spanning more than two decades, making it an ideal data source. However, we note that the primary contribution of our project remains the LLM training and evaluation (Sections 4–5). This section is not meant to provide a comprehensive analysis of policies in the Princeton-Leuven dataset. Rather, we intend the following examples to inspire future work using our LLM annotation method and the annotated policies we provide. **All 456 policies we annotated for the following analyses are publicly available at at the GitHub repository for this paper.**[8]

### 7.1 Longitudinal Privacy Policy Analysis

First, we chose 8 prominent companies and organizations[9] representing a variety of sectors, including "big tech," news, entertainment, finance, and government. This variety is useful because each sector has a unique approach to data collection, user engagement, and compliance with privacy regulations. Furthermore, these websites have undergone varying levels of public scrutiny. For instance, while Facebook and Google have faced major privacy debates, leading to numerous changes in their privacy policies, entities like *nsf.gov* operate under distinct governmental standards. This list also highlights geographic diversity concerning headquarters and user base, with some organizations primarily serving U.S. audiences while others have a global reach, necessitating compliance with various international privacy laws per the Brussels effect [13].

For each of these companies and organizations, we used our model to annotate one privacy policy from every year that the company or organization appears in the dataset (128 policies total).

[8]https://github.com/JakeC007/Automated_GKC-CI_Privacy_Policy_Annotations
[9]Facebook, The New York Times, GitHub, Buzzfeed, Google, Bank of America, Electronic Frontier Foundation (EFF), and the National Science Foundation (NSF)

The number of parameters in the policies of each company or organization over time are presented in Appendix I (Figure 17 and Tables 6–7).

The results provide insight into the evolution of privacy policies. For instance, we notice a generally increasing trend in the number of GKC-CI parameters included in privacy policies over time. As specific examples, the privacy policies of Buzzfeed and GitHub described fewer than 60 GKC-CI parameters in their policies from 2008-2010, but now describe over 400 or 500 parameters, respectively. The increase in parameters in the GitHub policy from 2016 to 2019 corresponds to the acquisition of GitHub by Microsoft.

The EFF and the NSF show similar, if less dramatic, increases in the number of parameters over time. This trend mirrors previously documented increases in average privacy policy length from 1996 to 2021 [81], providing a sanity check for our method – we expect longer privacy policies to include more details about information transfers. Indeed, the New York Times privacy policy underwent a dramatic *decrease* in the number of GKC-CI parameters in 2006, corresponding to an approximately 80% decrease in the length of the policy (23736 to 4912 words). The number of parameters then increased to above its previous maximum in 2011 when the policy length increased to 33616 words.

We also notice that although parameter counts are generally increasing and roughly track total policy length, the relative percentages of different parameters remains nearly consistent within each company's policy. This suggests that although companies are adding additional details to descriptions of data transfers in their privacy policies, these additions are not broadly skewed toward specific parameters. To understand the importance of this result, consider some counterfactual examples: If the relative percentage of *aim* parameters were to have increased, it would indicate that organizations are increasingly using privacy policies to inform *why* information is being collected over *what* information is being collected. If the relative percentage of *attribute* parameters were to have increased, it might indicate that organizations are collecting more data types per information transfer. While we don't see either of these trends for these 8 companies, we examine the variance of parameter types in the privacy policies of a larger set of companies in Section 7.2.

A closer look at the GitHub privacy policies from 2016 and 2019 demonstrates how GKC-CI annotation can be used to automatically identify privacy policy updates of interest. Specifically, the 2019 policy includes more *aim* parameters describing new ways that the site may use collected information, such as "to make recommendations for you, such as to suggest projects you may want to follow or contribute to" and "to determine your coding interests." Noticing a substantial increase in the number of *aims* and focusing directly on those new parameters allows for quick assessment of changes in information use. Similarly, automatically identifying new or changed *attributes* would allow quick assessment of new types of information collected.

Privacy policy updates happen regularly on a vast number of online services, and it is difficult for experts, to say nothing of average users, to identify which updates are salient for particular privacy concerns. We anticipate that our LLM annotation approach could be used in the future to automatically sort updates by parameter type and, combined with a visual interface, faster review of privacy policy updates.

## 7.2 Cross-Industry Privacy Policy Analysis

We next used our fine-tuned LLM to annotate the most recent privacy policies within the Tranco top 300 [53] websites that are in the Princeton-Leuven corpus (164 policies total).

In each of the following analyses, we highlight extreme examples from across these 164 policies to demonstrate how annotation at scale facilitates directed data exploration. Previous work has shown that detailed analysis of individual annotated policies using the CI framework can identify specific ambiguities and normative shortcomings [67]. While deep analysis of individual policies is out of scope for this paper, the following paragraphs show how GKC-CI annotation can be used to identify policies that might be worth such detailed exploration in future work.

*Parameter type variance.* We first calculated the variance in the percentages of individual parameter types across all annotated parameters in each policy. Previous work using CI annotation emphasized that descriptions of information transfers that are missing specific parameter types or that included substantially more specific parameter types ("parameter bloating") lead to ambiguities about the actual data handling practices of the organization [67]. Since policies with a greater variance in the percentages of individual parameter types are more likely to exhibit these issues, we rank our annotated policies by this metric.

Figure 6 shows the fifteen policies with the highest variance of parameter type percentages. This includes a policy from *apache.org* with relatively few *aim* parameters and a policy from *mozilla.org* with relatively few *attribute* parameters. A quick review of the *mozilla.org* policy shows why – it defines all data covered by the policy in one sentence: "For us, 'personal information' means information which identifies you, like your name or email address." The policy does not say which particular forms of personal information are connected to the descriptions of information use in the rest of the policy, leaving it ambiguous whether e.g., what specific customer data might be shared with other entities for "processing or providing products and services." While this type of high-level analysis doesn't necessarily imply the existence of policy ambiguities, it suggests that policies with high parameter type variance are promising candidates for a detailed evaluation through the lens of the GKC-CI framework. The parameter percentages for all 164 privacy policies are provided in Tables 8–9 in Appendix J.

*Parameter to sentence ratio.* We next calculated the ratio of annotated GKC-CI parameters to the number of sentences in each policy. This provides a metric of the "density" of information transfer descriptions in the policy. Figure 7 shows these data for the 15 privacy policies with the highest ratio of annotated parameters to sentences. The ratios for all 164 privacy policies are provided in Tables 10–11 in Appendix K. The top 15 policies include those from content distribution networks (*b-cdn.net*), web component frameworks (*ampproject.org*), Facebook, Microsoft (*bing.com*), Apple (*icloud.com*), Zoom, and social media websites (*linkedin.com*, *snapchat.com*, *t.co*, *tumblr.com*), among others.

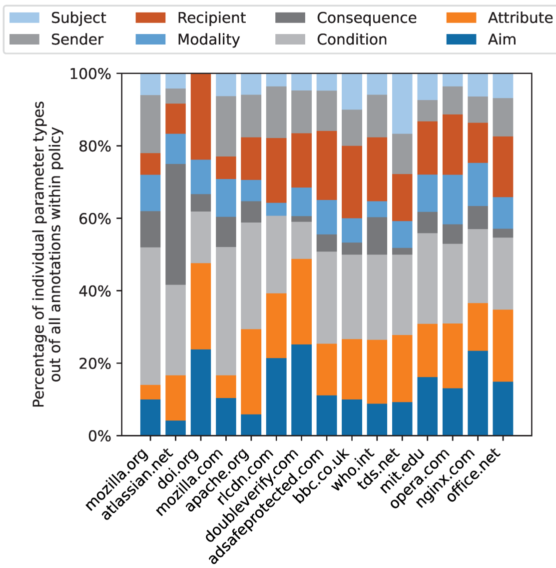


**Figure 6: The 15 privacy policies with the highest variance in the percentage of individual parameter types across all parameters annotated in the policy.**

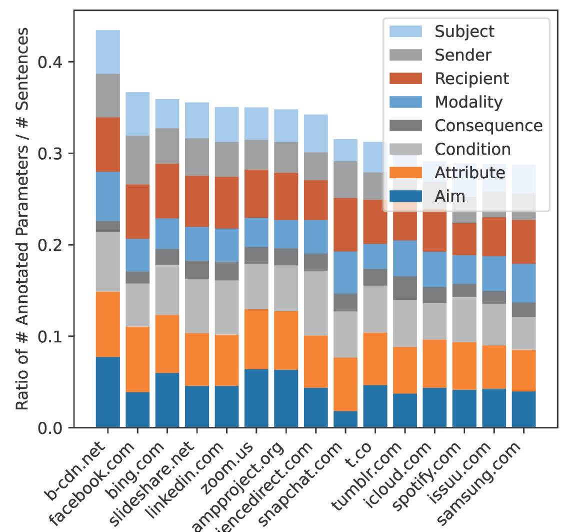


**Figure 7: The 15 privacy policies with the highest ratio of GKC-CI parameters to sentences out of all 164 of the Tranco top 300 websites in the Princeton-Leuven corpus.**

While these policies may exhibit parameter bloating issues due to the density of parameters, they may also be good examples of policies providing meaningful details about data handling practices. The first few policies on this list provide a microcosm of this variety. The *b-cdn.net* policy is very minimal, but each sentence is a short, to-the-point data handling description. The Facebook policy is longer, but an earlier 2018 version was identified in [67] as having parameter bloat. Either way, directing future in-depth investigations toward policies with high parameter density would provide examples of GKC-CI information flow descriptions for case studies for teaching [5] or iteration on the GKC-CI framework.

*Website popularity.* As a comparison against the 164 popular websites from the previous analyses, we next annotated the 164 policies from the Princeton-Leuven corpus with the *lowest* Tranco rankings (999981 to 991993). Figure 8 compares the distributions of total parameter counts and parameter to sentence ratios between these sets of websites.

The distributions of total parameter counts are significantly different ($p < 0.005$, Mann-Whitney U test), with the more popular websites having more GKC-CI parameters in their privacy policies on average than the less popular websites (mean 412 versus 130). This result makes sense, as more popular websites are under more scrutiny about their handling of user information and therefore include more information about data practices in their privacy policies. However, the distributions of parameter to sentence ratios are not significantly different ($p > 0.05$, Mann-Whitney U test), indicating that the more popular websites are generally providing more details by adding to policy length, rather than by increasing the density of information flow descriptions.

We anticipate that automated GKC-CI parameter annotation will enable more detailed statistical analyses across privacy policies in the future. For example, one could track the changing correlation between *condition* parameters in privacy policies and the text of data privacy regulations as both policies and regulation are updated – identifying outliers for scrutiny.

### 7.3 GKC-CI Annotation Visualizer

We have developed a visualizer tool designed to enhance the interpretability of our LLM's privacy policy annotations. This tool was created to empower researchers and others using our approach to better understand the complex data flows and governing principles that underpin privacy policies by visually distinguishing between different parameter tags. By highlighting various aspects of the policy in different colors, the visualizer makes the described information flows more accessible, allowing users to quickly grasp how their data may be used and protected under the terms of the policy.

The tool functions as a local Python script that processes both the raw text of the privacy policy and the output log from the LLM. It begins by matching specific text segments from the policy with those used in the LLM prompts. The visualizer then highlights these segments in different colors based on the associated annotations, making it easy for users to see which parts of the text correspond to specific GKC-CI parameters. This highlighted text is presented in a user-friendly GUI window, where users can scroll through and interact with the augmented policy. Additionally, it generates an output file that explains the highlighted segments, helping users understand the implications of the annotations. Note that if the LLM's output has not match in the text we consider the output incorrect and do not include these instances in the visualization.

To illustrate the capabilities of the visualizer, we applied it to two versions of Facebook's privacy policy from 2015 and 2019.

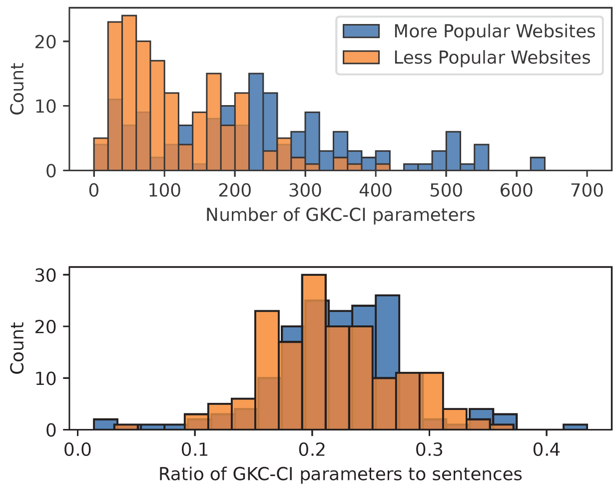


**Figure 8: Comparison of all 164 websites in the Tranco top 300 versus the 164 websites lowest on the Tranco rankings in the Princeton-Leuven corpus. *Top:* Distribution of the raw number of GKC-CI parameters annotated by our model. Outliers with** $> 700$ **parameters (13 policies for the more popular websites and 1 policy for the less popular websites) are omitted. *Bottom:* Distribution of the ratio of annotated GKC-CI parameters to sentence count.**

Comparing visualized excerpts of these versions (Appendix L) effectively highlights how data practices and governing principles have evolved. For instance, there are more GKC-CI parameters in the 2019 excerpt than the 2015 excerpt, particularly more *aims* and *consequences*. The new policy does a better job informing users about the repercussions of the described data collection. The visualizer's color-coded presentation makes it easier to understand how the policy has changed and will facilitate future research using GKC-CI annotation.

## 8 Future Directions

As demonstrated in Section 7, the ability to accurately annotate privacy policies with GKC-CI parameter tags enables a variety of previously infeasible analyses. We are excited about the potential for the GKC and CI community to use our method to facilitate advances at the intersection of machine learning and privacy.

A common question posed by CI researchers is whether information flows (composed of GKC-CI parameters) are *appropriate* in their respective contexts. This is a core element of the CI framework, which understands privacy breaches as inappropriate information flows that violate contextual norms [48]. There are many ways to discover the norms of a particular context, including surveys [6, 7, 70], interviews [42], focus groups [66], and textual analysis [10], among others. We propose using LLM annotation as the first step in a multi-method pipeline for this task: 1) a researcher extracts GKC-CI parameters from segments of an organization's privacy policy using LLM annotation, 2) the parameters are piped into a survey template about contextual appropriateness (like that in [6]), 3) crowdsourced survey responses from relevant community members indicate (mis)alignment with existing norms. This entire process could be automated from the perspective of the researchers, enabling much larger scale CI-based audits of data handling practices across many organizations than previous possible.

We also believe that GKC-CI annotation with LLMs need not be limited to privacy policies. Many documents, from white papers to media reports, describe transfers of information. We expect that fine-tuned LLMs will be able to identify GKC-CI parameters as accurately in those documents as in privacy policies. This opens an new set of applications that we hope the community will pursue.

## 9 Conclusion

This paper demonstrates that high-accuracy annotation of contextual integrity (CI) and governing knowledge commons (GKC) parameters in privacy policies can be achieved using LLMs. We ultimately find that `GPT 3.5TPE_25ep` had the best performance, with an accuracy of 90.65% for exact string matches. While we find that proprietary LLMs outperformed open-source models, we report some valuable findings for researchers interested in performing LLM application studies. Namely, 1) that LLM size must be considered in context to model family; smaller, aligned models are the most economical choice, and 2) that library defaults are likely to introduce confounds and should be checked.

We demonstrate the usefulness of our fine-tuned model by annotating 456 privacy policies from the Princeton-Leuven Longitudinal Corpus of Privacy Policies [1]. We show that large-scale GKC-CI annotation can be an effective tool for data exploration, highlighting changes in parameter frequency over time, policies with relatively high variances across parameter type percentages, and policies with relatively high ratios of parameters to sentences. This facilitates automated review of privacy policy updates to identify meaningful changes in information flow descriptions, potentially with normative implications. We have made our model training code, training and testing data, annotation visualizer, and all annotated policies publicly available[10] in the hope that this work motivates future use of GKC-CI parameter annotation.

## Acknowledgments

We thank research assistants Sophia Goffe and Wael Mohamed for their contributions. This research was funded by the Colgate University Faculty Research Council and the Colgate University Department of Computer Science. Some of the results presented in this paper were obtained using the Chameleon testbed [40] supported by the National Science Foundation.

## References

[1] 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*. 2165–2176.
[2] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating internal privacy policy contradictions on Google Play. In *28th USENIX Security Symposium (USENIX Security 19)*. 585–602.
[3] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words:

[10] https://github.com/JakeC007/Automated_GKC-CI_Privacy_Policy_Annotations

Entity-sensitive privacy policy and data flow analysis with PoliCheck. In *29th USENIX Security Symposium (USENIX Security 20)*. 985–1002.
[4] Julio Angulo, Simone Fischer-Hübner, Erik Wästlund, and Tobias Pulls. 2012. Towards usable privacy policy display and management. *Information Management & Computer Security* (2012).
[5] Noah Apthorpe. 2021. Practical assignments for teaching contextual integrity. In *3rd Annual Symposium on Applications of Contextual Integrity*.
[6] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–23.
[7] Noah Apthorpe, Sarah Varghese, and Nick Feamster. 2019. Evaluating the contextual integrity of privacy regulation: Parents' IoT toy privacy norms versus COPPA. In *28th USENIX Security Symposium (USENIX Security 19)*. 123–140.
[8] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. 2020. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*. 1943–1954.
[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
[10] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. 2006. Privacy and contextual integrity: Framework and applications. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 15–pp.
[11] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).
[12] Sebastian Benthall. 2022. Integrating differential privacy and contextual integrity. USENIX Association, Santa Clara, CA.
[13] Anu Bradford. 2012. The Brussels effect. *Northwestern University Law Review* 107 (2012), 1.
[14] Karoline Brehm and Yan Shvartzhnaider. 2023. Understanding privacy in virtual reality classrooms: A contextual integrity perspective. *IEEE Security & Privacy* (2023).
[15] Carolyn A Brodie, Clare-Marie Karat, and John Karat. 2006. An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. In *Proceedings of the Second Symposium on Usable Privacy and Security*. 8–19.
[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
[17] Cengage. 2022. https://www.cengagegroup.com/privacy/notice/.
[18] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
[19] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
[20] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
[21] Crowdmark. 2022. https://crowdmark.com/privacy/.
[22] Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. LAiW: A Chinese legal large language models benchmark (a technical report). *arXiv preprint arXiv:2310.05620* (2023).
[23] Anupam Datta, Jeremiah Blocki, Nicolas Christin, Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kaynar, and Arunesh Sinha. 2011. Understanding and protecting privacy: Formal semantics and principled audit mechanisms. In *International Conference on Information Systems Security*. Springer, 1–27.
[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[25] Dropbox. 2022. https://www.dropbox.com/privacy.
[26] Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L Littman. 2021. Brittle AI, causal confusion, and bad mental models: Challenges and successes in the XAI program. *arXiv preprint arXiv:2106.05506* (2021).
[27] Facebook. 2022. https://m.facebook.com/privacy/explanation/.
[28] Facebook Research. 2023. Llama Recipes. https://github.com/facebookresearch/llama-recipes.
[29] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. LawBench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289* (2023).
[30] Brett M Frischmann, Michael J Madison, and Katherine Jo Strandburg. 2014. *Governing knowledge commons*. Oxford University Press.
[31] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2021).
[32] Gradescope. 2022. https://www.gradescope.com/privacy.
[33] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*. 531–548.
[34] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).
[35] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2020).
[36] Honorlock. 2022. https://honorlock.com/wp-content/uploads/2021/05/May2021_Honorlock_App_Privacy_Policy.docx.pdf.
[37] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
[38] Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 471–478.
[39] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
[40] Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.
[41] Kultura2022. 2022. https://corp.kaltura.com/legal/privacy/privacy-policy/.
[42] Priya C Kumar, Michael Zimmer, and Jessica Vitak. 2024. A roadmap for applying the contextual integrity framework in qualitative privacy research. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.
[43] LinkedIn. 2022. https://www.linkedin.com/legal/privacy-policy.
[44] Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A Smith. 2014. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 884–894.
[45] Matlab. 2022. https://www.mathworks.com/company/aboutus/policies_statements/privacy-policy.html.
[46] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4 (2008), 543.
[47] Niantic. 2022. https://nianticlabs.com/privacy/en/.
[48] Helen Nissenbaum. 2009. *Privacy in Context*. Stanford University Press.
[49] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[50] Packback. 2022. https://www.packback.co/site/privacy/.
[51] Panopto. 2022. https://www.panopto.com/privacy/.
[52] Alfredo J Perez, Sherali Zeadally, and Jonathan Cochran. 2018. A review and an empirical analysis of privacy policy and notices for consumer Internet of things. *Security and Privacy* 1, 3 (2018), e15.
[53] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2018. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156* (2018).
[54] Ofir Press, Noah A Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409* (2022).

[55] Proctorio. 2022. https://proctorio.com/privacy.
[56] Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. https://api.semanticscholar.org/CorpusID:49313245
[57] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://api.semanticscholar.org/CorpusID:160025533
[58] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841* (2019).
[59] Joel R Reidenberg, Jaspreet Bhatia, Travis D Breaux, and Thomas B Norton. 2016. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies* 45, S2 (2016), S163–S190.
[60] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal* 30 (2015), 39.
[61] Sean Robertson. 2024. https://pytorch.org/tutorials/intermediate/char_rnn_generation_tutorial.html
[62] Manuel Rudolph, Denis Feth, and Svenja Polst. 2018. Why users ignore privacy policies–a survey and intention model for explaining user privacy behavior. In *International Conference on Human-Computer Interaction*. Springer, 587–598.
[63] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The usable privacy policy project. In *Technical Report, CMU-ISR-13-119*. Carnegie Mellon University.
[64] Madelyn Sanfilippo, Brett Frischmann, and Katherine Standburg. 2018. Privacy as commons: Case evaluation through the governing knowledge commons framework. *Journal of Information Policy* 8, 1 (2018), 116–166.
[65] Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? *arXiv preprint arXiv:2103.08493* (2021).
[66] Gwen Shaffer. 2021. Applying a contextual integrity framework to privacy policies for smart technologies. *Journal of Information Policy* 11 (2021), 222–265.
[67] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2019. Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 162–170.
[68] Yan Shvartzshnaider, Zvonimir Pavlinovic, Ananth Balashankar, Thomas Wies, Lakshminarayanan Subramanian, Helen Nissenbaum, and Prateek Mittal. 2019. Vaccine: Using contextual integrity for data leakage detection. In *The World Wide Web Conference*. 1702–1712.
[69] Yan Shvartzshnaider, Madelyn Rose Sanfilippo, and Noah Apthorpe. 2022. GKC-CI: A unifying framework for contextual norms and information governance. *Journal of the Association for Information Science and Technology* (2022).
[70] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning privacy expectations by crowdsourcing contextual informational norms. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. 209–218.
[71] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 102–107.
[72] Peter Story, Sebastian Zimmeck, Abhilasha Ravichander, Daniel Smullen, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Natural language processing for mobile app privacy compliance. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*, Vol. 2. 4.
[73] Stripe. 2022. https://stripe.com/privacy.
[74] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. 2023. PolicyGPT: Automated analysis of privacy policies with large language models. *arXiv preprint arXiv:2309.10238* (2023).
[75] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. 15–21.
[76] The New York Times. 2022. https://www.nytimes.com/privacy/privacy-policy.
[77] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[78] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[79] Turnitin. 2022. https://help.turnitin.com/Privacy_and_Security/Privacy_and_Security.htm#Privacy_Policy.
[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[81] Isabel Wagner. 2022. Privacy policies across the ages: Content and readability of privacy policies 1996–2021. *arXiv preprint arXiv:2201.08739* (2022).
[82] Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2023. Reformulating domain adaptation of large language models as adapt-retrieve-revise. *arXiv preprint arXiv:2310.03328* (2023).
[83] Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247* (2022).
[84] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2022).
[85] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. 2015. Android permissions remystified: A field study on contextual integrity. In *24th USENIX Security Symposium (USENIX Security 15)*. 499–514.
[86] 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, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1330–1340.
[87] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, et al. 2018. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web (TWEB)* 13, 1 (2018), 1–29.
[88] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing annotations for websites' privacy policies: Can it really work?. In *Proceedings of the 25th International Conference on World Wide Web*. 133–143.
[89] Stephanie Winkler and Sherali Zeadally. 2016. Privacy policy analysis of popular web platforms. *IEEE Technology and Society Magazine* 35, 2 (2016), 75–85.
[90] Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. DISC-LawLLM: Fine-tuning Large Language Models for intelligent legal services. *arXiv preprint arXiv:2309.11325* (2023).
[91] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
[92] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 66–86.

# A Ground Truth Details

| Company | Word Count | Sender | Subject | Recipient | Attribute | Aim | Condition | Modality | Consequence |
|---|---|---|---|---|---|---|---|---|---|
| Cengage [17] | 3340 | 12 | 15 | 34 | 44 | 64 | 82 | 30 | 0 |
| Crowdmark [21] | 3216 | 12 | 28 | 37 | 42 | 80 | 92 | 42 | 6 |
| Dropbox [25] | 2485 | 10 | 9 | 20 | 26 | 36 | 30 | 10 | 4 |
| Facebook [27] | 4151 | 40 | 48 | 74 | 113 | 78 | 84 | 42 | 0 |
| Gradescope [32] | 11431 | 18 | 20 | 39 | 69 | 104 | 152 | 28 | 2 |
| Honorlock [36] | 1199 | 10 | 9 | 11 | 22 | 18 | 32 | 22 | 2 |
| Kultura [41] | 6255 | 15 | 29 | 54 | 63 | 96 | 164 | 52 | 4 |
| LinkedIn [43] | 6298 | 37 | 58 | 80 | 111 | 110 | 174 | 22 | 0 |
| Matlab [45] | 5580 | 27 | 27 | 61 | 85 | 98 | 150 | 44 | 4 |
| Niantic [47] | 5539 | 27 | 33 | 44 | 63 | 92 | 94 | 16 | 2 |
| NYTimes [76] | 5000 | 12 | 25 | 41 | 50 | 50 | 82 | 18 | 2 |
| Packback [50] | 4444 | 11 | 14 | 25 | 35 | 40 | 94 | 18 | 4 |
| Panopto [51] | 4167 | 17 | 16 | 28 | 34 | 62 | 82 | 42 | 2 |
| Proctorio [55] | 9353 | 28 | 24 | 61 | 89 | 124 | 140 | 54 | 2 |
| Stripe [73] | 7460 | 38 | 48 | 73 | 96 | 110 | 122 | 40 | 4 |
| Turnitin [79] | 10220 | 15 | 24 | 24 | 52 | 94 | 90 | 18 | 4 |

**Table 3: Number of labeled parameters in ground-truth GKC-CI annotations of 16 privacy policies (rows) from popular websites and e-learning services. All of the labels from the manually-annotated privacy policies were combined together to create the ground-truth dataset used for evaluating all models. There is no difference in which labels were used for training/testing across different models.**

# B Brat Annotation Legend

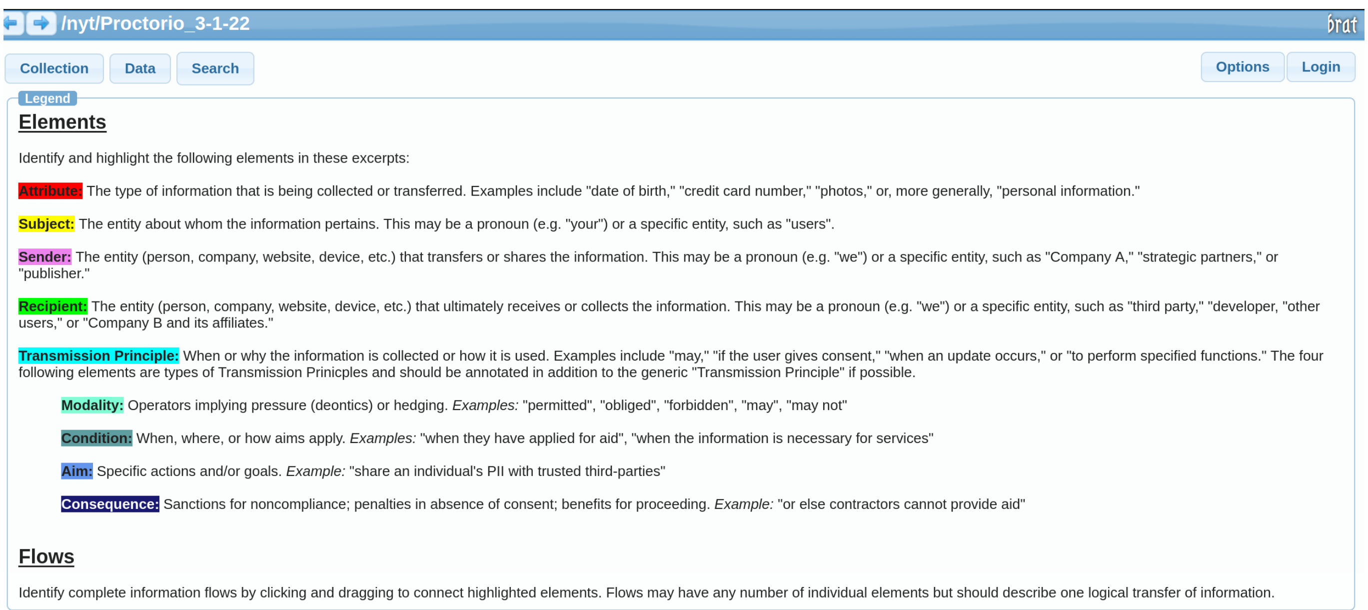


**Figure 9: Legend in the customized brat annotation tool [71] for expert annotators to use as reference.**

## C Performance On Sentence Versus Paragraph-Delineated Data As Epochs Vary

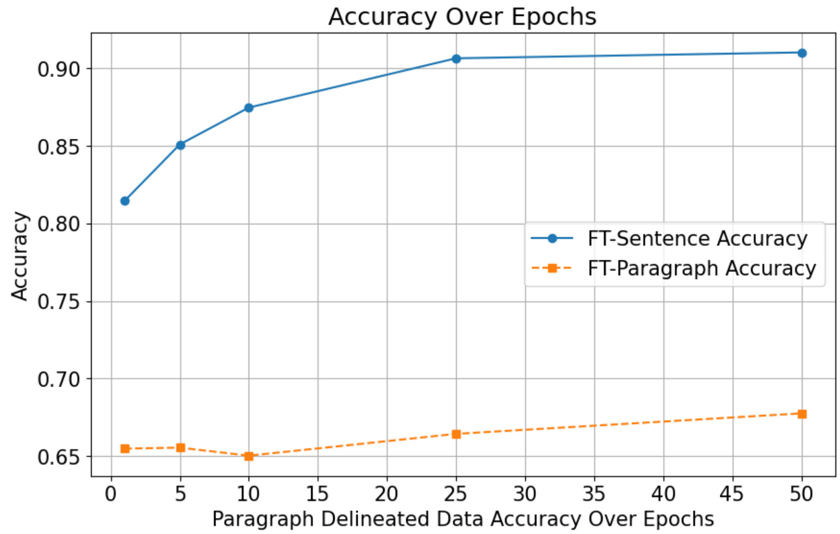


**Figure 10: The solid blue line ("FT-Sentence Accuracy") is the test-set accuracy of `GPT 3.5TPE` trained on sentence-delineated data at 1, 5, 10, 25, and 50 epochs. The dotted orange line ("FT-Paragraph Accuracy") is the test-set accuracy of `GPT 3.5TPE` trained on paragraph-delineated data at 1, 5, 10, 25, and 50 epochs. An accurate annotation is one that is an exact string match.**

## D Baseline Non-fine-tuned performance With N-Shot Learning as N varies

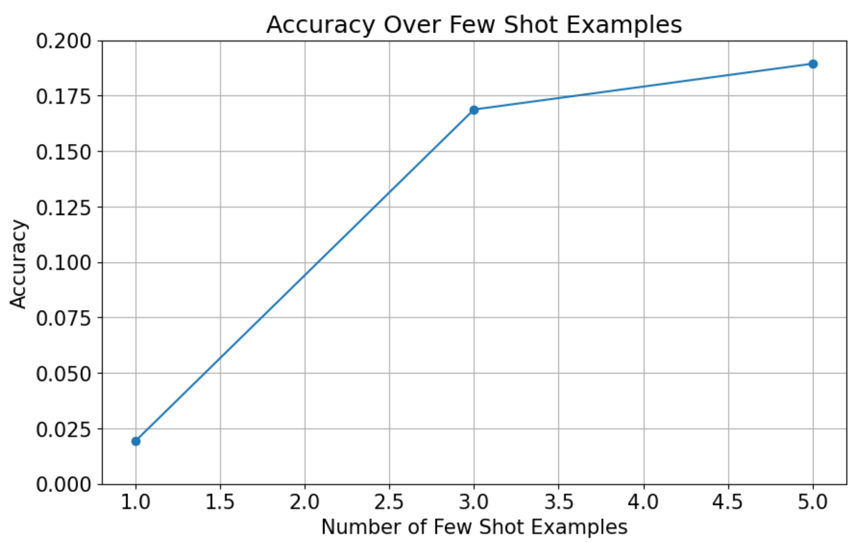


**Figure 11: Test-set accuracy of non-fine-tuned `GPT 3.5TPE` with simple prompting and 1, 3 and 5 few shot examples. An accurate annotation is one that is an exact string match. Note the maximum accuracy of <20%.**

## E Expanded Model Names

| Model Name | Model Family | Model Size | BOS, EOS Tokens Added? | Training Objective | Epochs |
|---|---|---|---|---|---|
| RNN_boseos_10ep | N/A: Recurrent Neural Network | 67M | Yes | CLM | 10 |
| llama2-7b_boseos_10ep | llama2 | 7B | No | CLM | 10 |
| gpt3,5_tpe_10ep | GPT 3.5 Turbo | Propetary | Propetary | CLM | 10 |
| gpt3,5_tg_10ep | GPT 3.5 Turbo | Propetary | Propetary | CLM | 10 |
| gpt3,5_t2s_10ep | GPT 3.5 Turbo | Propetary | Propetary | CLM | 10 |
| gpt2_xl_boseos_10ep | GPT 2 | XL (1.5B) | Yes | CLM | 10 |
| gpt2_xl_10ep | GPT 2 | XL (1.5B) | No | CLM | 10 |
| gpt2_boseos_10ep | GPT 2 | Base (124M) | Yes | CLM | 10 |
| gpt2_10ep | GPT 2 | Base (124M) | No | CLM | 10 |
| flan-t5_seq2seq_boseos_1ep | Flan-T5 | Base (248M) | Yes | MLM | 1 |
| flan-t5_seq2seq_1ep | Flan-T5 | Base (248M) | No | MLM | 1 |
| flan-t5_large_seq2seq_1ep | Flan-T5 | Large (783M) | No | MLM | 1 |
| flan-t5_large_seq2seq_boseos_1ep | Flan-T5 | Large (783M) | Yes | MLM | 1 |
| flan-t5_large_seq2seq_1ep | Flan-T5 | Large (783M) | No | MLM | 1 |

**Table 4: Model names as they correspond to their specific training/fine-tuning interventions.**

## F Open-Source Model Performance as Epochs Vary

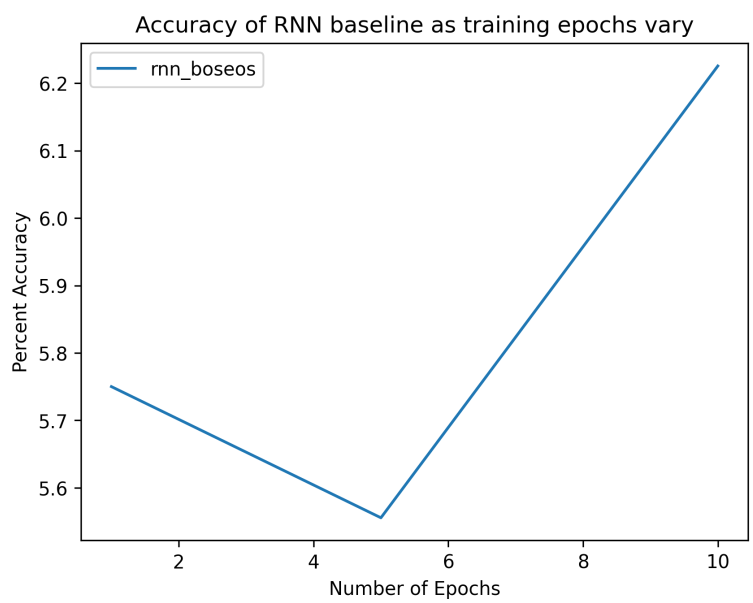


**Figure 12: Test-set accuracy of our RNN at 1, 5, and 10 epochs. An accurate annotation is one that is an exact string match.**

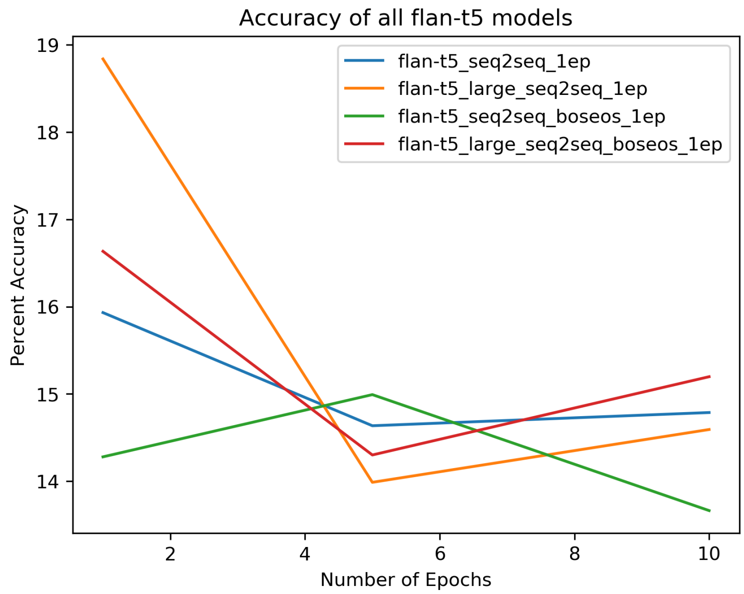


**Figure 13: Test-set accuracy of Flan-T5 models at 1, 5, and 10 epochs. An accurate annotation is one that is an exact string match.**

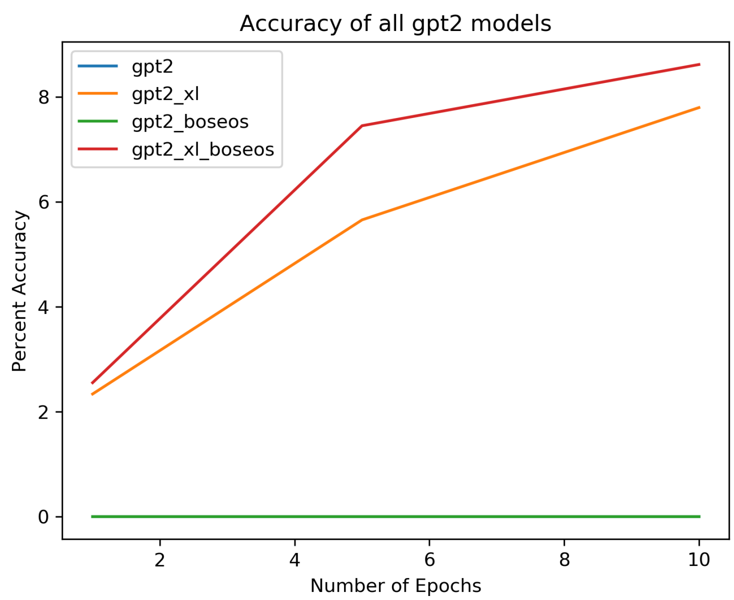


**Figure 14: Test-set accuracy of GPT-2 models at 1, 5, and 10 epochs. An accurate annotation is one that is an exact string match.**

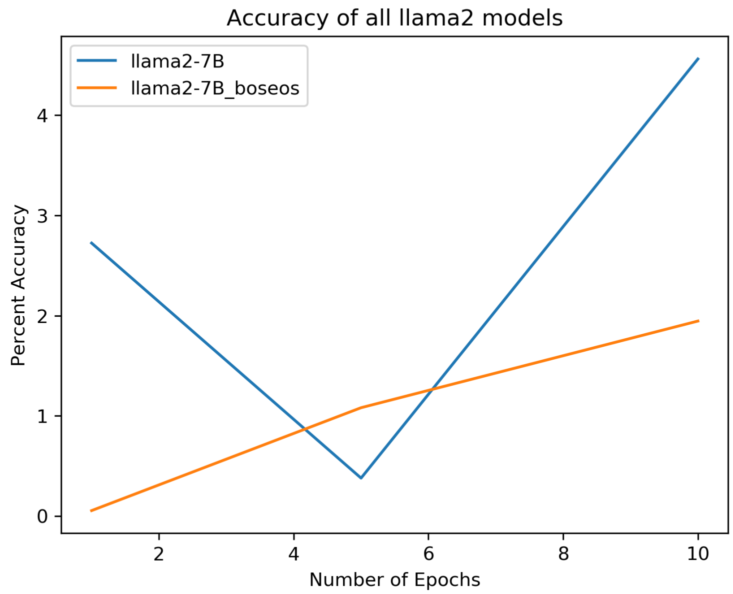


**Figure 15: Test-set accuracy of Llama2 models at 1, 5, and 10 epochs. An accurate annotation is one that is an exact string match.**

# G Performance Breakdown For The Three Top Performing Models

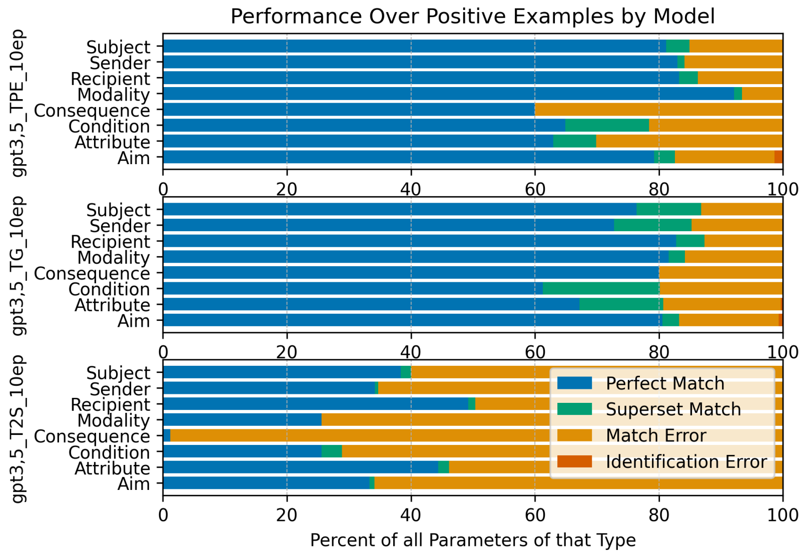


**Figure 16: A comparison of our top-performing models on positive examples by GKC-CI parameter. `GPT3,5_TG_10ep` refers the generic GPT-3.5 Turbo model, `GPT3,5_TPE_10ep` refers to the prompt-engineered version of GPT-3.5 Turbo, and `GPT3,5_T2S_10ep` refers to the two model GPT-3.5 Turbo system. All models were trained for 10 epochs. Note that there are only 5 "Consequence" parameters leading to greater variance on the corresponding bars.**

# H Codebook

| Code | Description |
|---|---|
| **Completion Errors** | |
| Completion Is Wrong | Completion is outright incorrect failing to provide the accurate answer. |
| Meaningful Subset | Completion partially captures the correct response but falls short of completeness. |
| Completion Over-labeled | Completion includes correct answers but erroneously incorporates nearby words into the parameter tag. |
| **Expert Labeling Errors** | |
| Expert Labels Is Wrong | Expert label itself is incorrect. |
| Expansive Ground Truth | Expert label is correct but overly broad and the completion offers a more precise response. |
| Partial Ground Truth | Expert label misses a portion of the correct label, but the completion captures it accurately. |
| **Semantic Equivalence** | |
| Semantic Equivalence | Completion and the ground truth label differ in wording but convey equivalent semantic meanings. |

**Table 5: Codebook for qualitative error analysis. Parent codes in bold.**

# I Longitudinal Data

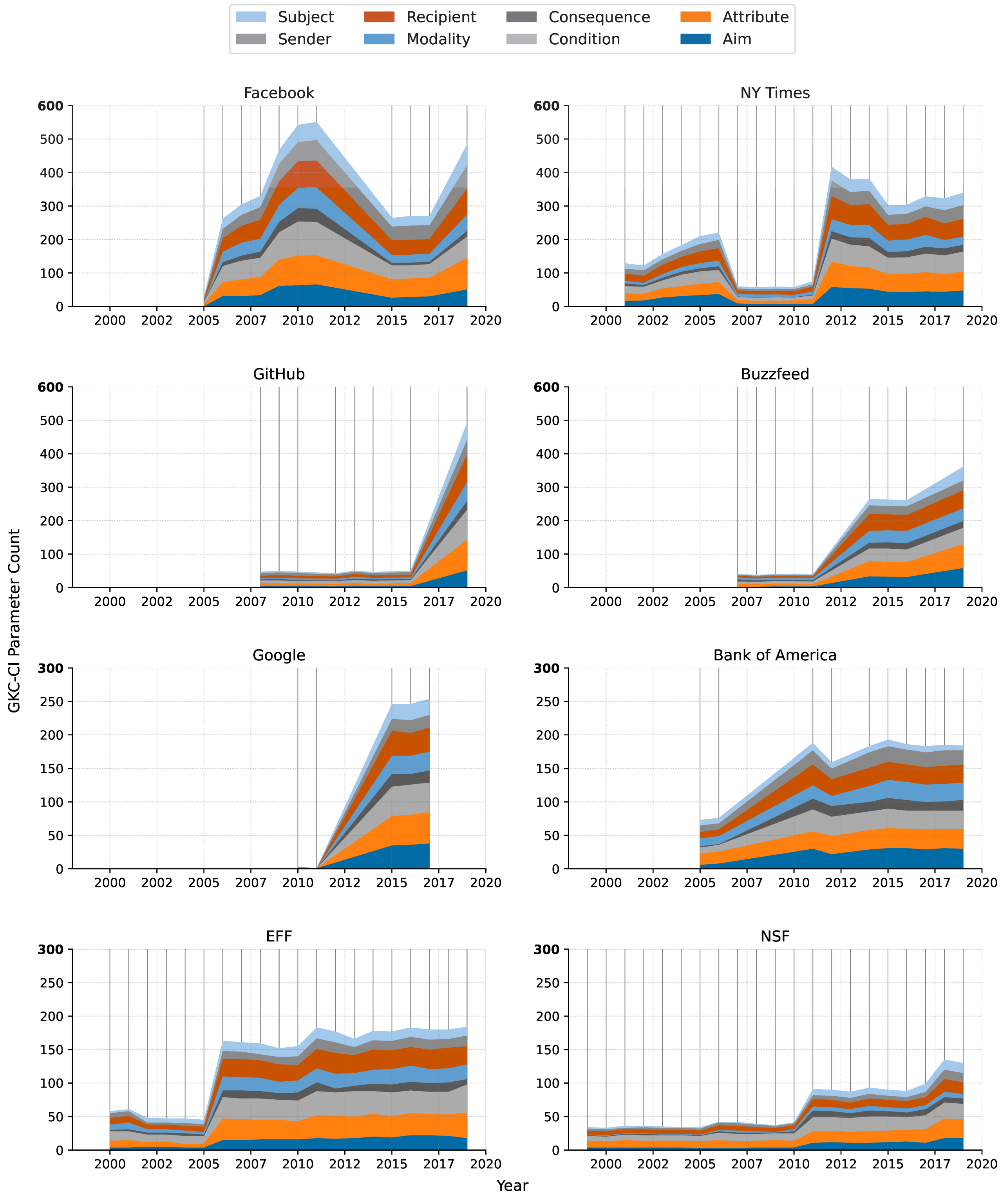


**Figure 17: Number of annotated GKC-CI parameters in the privacy policies of 8 prominent companies over time. Policies from the Princeton-Leuven Longitudinal Corpus of Privacy Policies [1]. Bold grid lines indicate which years' policies were in the corpus and were annotated. The exact parameter counts displayed in this figure appear in Tables 6–7 in Appendix I.**

| Company | Year | Aim | Attribute | Condition | Consequence | Modality | Recipient | Sender | Subject | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Facebook | 2005 | 1 | 7 | 5 | 0 | 1 | 6 | 4 | 5 | 29 |
| Facebook | 2006 | 31 | 42 | 48 | 11 | 32 | 40 | 28 | 30 | 262 |
| Facebook | 2007 | 31 | 50 | 56 | 15 | 39 | 50 | 33 | 31 | 305 |
| Facebook | 2008 | 34 | 54 | 58 | 18 | 39 | 56 | 38 | 33 | 330 |
| Facebook | 2009 | 62 | 78 | 82 | 31 | 51 | 70 | 52 | 41 | 467 |
| Facebook | 2010 | 63 | 89 | 102 | 40 | 61 | 79 | 57 | 51 | 542 |
| Facebook | 2011 | 66 | 87 | 100 | 39 | 64 | 80 | 61 | 54 | 551 |
| Facebook | 2015 | 26 | 55 | 42 | 7 | 24 | 44 | 41 | 26 | 265 |
| Facebook | 2016 | 29 | 55 | 39 | 8 | 24 | 44 | 43 | 28 | 270 |
| Facebook | 2017 | 30 | 56 | 41 | 7 | 24 | 44 | 41 | 27 | 270 |
| Facebook | 2019 | 51 | 94 | 64 | 17 | 49 | 77 | 70 | 62 | 484 |
| NY Times | 2001 | 17 | 22 | 22 | 8 | 8 | 21 | 15 | 16 | 129 |
| NY Times | 2002 | 18 | 20 | 21 | 5 | 7 | 21 | 16 | 14 | 122 |
| NY Times | 2003 | 27 | 27 | 24 | 8 | 11 | 28 | 16 | 15 | 156 |
| NY Times | 2004 | 31 | 30 | 34 | 9 | 14 | 30 | 18 | 17 | 183 |
| NY Times | 2005 | 34 | 34 | 37 | 9 | 16 | 35 | 22 | 23 | 210 |
| NY Times | 2006 | 37 | 36 | 36 | 11 | 17 | 37 | 25 | 22 | 221 |
| NY Times | 2007 | 9 | 10 | 10 | 1 | 7 | 11 | 6 | 6 | 60 |
| NY Times | 2008 | 7 | 10 | 9 | 1 | 9 | 10 | 5 | 6 | 57 |
| NY Times | 2009 | 7 | 10 | 11 | 1 | 7 | 11 | 6 | 6 | 59 |
| NY Times | 2010 | 7 | 11 | 9 | 1 | 7 | 11 | 6 | 6 | 58 |
| NY Times | 2011 | 8 | 14 | 11 | 2 | 10 | 14 | 7 | 9 | 75 |
| NY Times | 2012 | 58 | 76 | 69 | 22 | 36 | 69 | 46 | 41 | 417 |
| NY Times | 2013 | 55 | 66 | 64 | 22 | 37 | 59 | 39 | 37 | 379 |
| NY Times | 2014 | 53 | 64 | 63 | 25 | 39 | 62 | 40 | 35 | 381 |
| NY Times | 2015 | 44 | 52 | 50 | 17 | 34 | 47 | 30 | 29 | 303 |
| NY Times | 2016 | 43 | 54 | 50 | 18 | 35 | 47 | 30 | 27 | 304 |
| NY Times | 2017 | 45 | 58 | 55 | 20 | 36 | 54 | 31 | 30 | 329 |
| NY Times | 2018 | 44 | 52 | 57 | 21 | 25 | 49 | 40 | 35 | 323 |
| NY Times | 2019 | 48 | 57 | 59 | 20 | 25 | 53 | 41 | 37 | 340 |
| GitHub | 2008 | 6 | 8 | 8 | 3 | 4 | 7 | 8 | 4 | 48 |
| GitHub | 2009 | 5 | 8 | 9 | 4 | 4 | 8 | 8 | 4 | 50 |
| GitHub | 2010 | 5 | 8 | 7 | 3 | 5 | 8 | 8 | 4 | 48 |
| GitHub | 2012 | 4 | 8 | 8 | 2 | 5 | 7 | 6 | 3 | 43 |
| GitHub | 2013 | 6 | 8 | 9 | 4 | 5 | 9 | 7 | 3 | 51 |
| GitHub | 2014 | 5 | 7 | 9 | 4 | 5 | 8 | 6 | 3 | 47 |
| GitHub | 2016 | 5 | 9 | 9 | 3 | 5 | 8 | 8 | 4 | 51 |
| GitHub | 2019 | 52 | 91 | 90 | 26 | 57 | 81 | 45 | 51 | 493 |
| Buzzfeed | 2007 | 3 | 8 | 8 | 6 | 6 | 5 | 3 | 2 | 41 |
| Buzzfeed | 2008 | 3 | 7 | 8 | 4 | 6 | 5 | 3 | 2 | 38 |
| Buzzfeed | 2009 | 4 | 8 | 8 | 4 | 6 | 5 | 4 | 2 | 41 |
| Buzzfeed | 2011 | 4 | 7 | 8 | 4 | 6 | 5 | 4 | 2 | 40 |
| Buzzfeed | 2014 | 34 | 45 | 38 | 17 | 36 | 50 | 26 | 18 | 264 |
| Buzzfeed | 2015 | 33 | 44 | 40 | 18 | 36 | 48 | 25 | 19 | 263 |
| Buzzfeed | 2016 | 32 | 45 | 37 | 19 | 37 | 47 | 26 | 18 | 261 |
| Buzzfeed | 2019 | 59 | 71 | 49 | 20 | 38 | 54 | 29 | 41 | 361 |
| Google | 2010 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| Google | 2011 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Google | 2015 | 35 | 44 | 44 | 19 | 27 | 37 | 18 | 22 | 246 |
| Google | 2016 | 36 | 45 | 45 | 16 | 27 | 34 | 19 | 24 | 246 |
| Google | 2017 | 38 | 47 | 44 | 18 | 28 | 36 | 19 | 24 | 254 |

**Table 6: Counts of annotated parameters in the privacy policies of 10 prominent companies and organizations over time *(continued on next page)*.**

| Company | Year | Aim | Attribute | Condition | Consequence | Modality | Recipient | Sender | Subject | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Bank of America | 2005 | 6 | 17 | 9 | 2 | 12 | 9 | 10 | 8 | 73 |
| Bank of America | 2006 | 8 | 18 | 10 | 1 | 12 | 10 | 9 | 8 | 76 |
| Bank of America | 2011 | 30 | 26 | 33 | 16 | 20 | 31 | 21 | 11 | 188 |
| Bank of America | 2012 | 22 | 27 | 29 | 16 | 15 | 25 | 16 | 9 | 159 |
| Bank of America | 2014 | 29 | 29 | 28 | 14 | 24 | 27 | 23 | 9 | 183 |
| Bank of America | 2015 | 31 | 30 | 29 | 16 | 27 | 27 | 23 | 10 | 193 |
| Bank of America | 2016 | 31 | 29 | 27 | 16 | 27 | 26 | 22 | 8 | 186 |
| Bank of America | 2017 | 29 | 30 | 28 | 13 | 26 | 26 | 22 | 9 | 183 |
| Bank of America | 2018 | 31 | 29 | 27 | 14 | 26 | 27 | 23 | 8 | 185 |
| Bank of America | 2019 | 30 | 29 | 28 | 16 | 26 | 27 | 21 | 7 | 184 |
| EFF | 2000 | 4 | 10 | 14 | 2 | 8 | 10 | 7 | 3 | 58 |
| EFF | 2001 | 4 | 11 | 13 | 3 | 10 | 10 | 7 | 3 | 61 |
| EFF | 2002 | 5 | 7 | 11 | 3 | 5 | 7 | 3 | 7 | 48 |
| EFF | 2003 | 5 | 8 | 10 | 3 | 5 | 7 | 3 | 6 | 47 |
| EFF | 2004 | 4 | 5 | 12 | 4 | 4 | 7 | 4 | 7 | 47 |
| EFF | 2005 | 4 | 6 | 11 | 2 | 4 | 8 | 4 | 6 | 45 |
| EFF | 2006 | 15 | 32 | 32 | 10 | 21 | 26 | 12 | 15 | 163 |
| EFF | 2007 | 15 | 31 | 31 | 12 | 20 | 27 | 11 | 14 | 161 |
| EFF | 2008 | 16 | 29 | 32 | 11 | 20 | 26 | 9 | 16 | 159 |
| EFF | 2009 | 16 | 29 | 30 | 10 | 17 | 26 | 11 | 13 | 152 |
| EFF | 2010 | 16 | 27 | 31 | 12 | 18 | 23 | 13 | 15 | 155 |
| EFF | 2011 | 18 | 34 | 36 | 13 | 21 | 29 | 16 | 16 | 183 |
| EFF | 2012 | 17 | 34 | 35 | 6 | 22 | 31 | 16 | 16 | 177 |
| EFF | 2013 | 18 | 32 | 38 | 8 | 19 | 27 | 12 | 12 | 166 |
| EFF | 2014 | 20 | 34 | 34 | 11 | 21 | 30 | 14 | 14 | 178 |
| EFF | 2015 | 19 | 32 | 35 | 12 | 23 | 28 | 14 | 14 | 177 |
| EFF | 2016 | 22 | 33 | 34 | 13 | 24 | 28 | 15 | 14 | 183 |
| EFF | 2017 | 22 | 32 | 33 | 13 | 21 | 29 | 15 | 15 | 180 |
| EFF | 2018 | 21 | 32 | 34 | 14 | 21 | 31 | 13 | 14 | 180 |
| EFF | 2019 | 18 | 38 | 42 | 8 | 22 | 27 | 16 | 13 | 184 |
| NSF | 1999 | 4 | 10 | 7 | 1 | 1 | 6 | 3 | 2 | 34 |
| NSF | 2000 | 4 | 9 | 7 | 1 | 1 | 6 | 3 | 2 | 33 |
| NSF | 2001 | 4 | 11 | 8 | 1 | 1 | 6 | 2 | 3 | 36 |
| NSF | 2002 | 4 | 10 | 8 | 2 | 1 | 6 | 3 | 2 | 36 |
| NSF | 2003 | 4 | 10 | 8 | 1 | 1 | 6 | 3 | 2 | 35 |
| NSF | 2004 | 4 | 10 | 8 | 1 | 1 | 6 | 3 | 1 | 34 |
| NSF | 2005 | 3 | 10 | 8 | 1 | 1 | 6 | 3 | 2 | 34 |
| NSF | 2006 | 3 | 12 | 11 | 2 | 2 | 7 | 4 | 1 | 42 |
| NSF | 2007 | 3 | 10 | 11 | 2 | 3 | 7 | 4 | 2 | 42 |
| NSF | 2008 | 4 | 10 | 10 | 2 | 2 | 6 | 4 | 1 | 39 |
| NSF | 2009 | 4 | 11 | 10 | 1 | 1 | 6 | 3 | 1 | 37 |
| NSF | 2010 | 4 | 10 | 11 | 3 | 3 | 6 | 3 | 1 | 41 |
| NSF | 2011 | 11 | 17 | 21 | 10 | 6 | 11 | 6 | 9 | 91 |
| NSF | 2012 | 12 | 17 | 20 | 9 | 6 | 11 | 6 | 9 | 90 |
| NSF | 2013 | 11 | 16 | 21 | 7 | 6 | 10 | 7 | 9 | 87 |
| NSF | 2014 | 11 | 18 | 21 | 9 | 7 | 11 | 7 | 9 | 93 |
| NSF | 2015 | 12 | 17 | 21 | 8 | 6 | 11 | 6 | 9 | 90 |
| NSF | 2016 | 13 | 17 | 19 | 7 | 6 | 11 | 6 | 9 | 88 |
| NSF | 2017 | 11 | 20 | 21 | 9 | 6 | 12 | 8 | 12 | 99 |
| NSF | 2018 | 18 | 29 | 24 | 8 | 8 | 19 | 14 | 15 | 135 |
| NSF | 2019 | 18 | 28 | 23 | 8 | 7 | 17 | 14 | 15 | 130 |

**Table 7: Counts of annotated parameters in the privacy policies of 10 prominent companies and organizations over time *(continued from previous page)***

# J Parameter Variance Data

| Website | % Aim | % Attribute | % Condition | % Consequence | % Modality | % Recipient | % Sender | % Subject | Variance |
|---|---|---|---|---|---|---|---|---|---|
| mozilla.org | 10.0 | 4.0 | 38.0 | 10.0 | 10.0 | 6.0 | 16.0 | 6.0 | 119.7 |
| atlassian.net | 4.2 | 12.5 | 25.0 | 33.3 | 8.3 | 8.3 | 4.2 | 4.2 | 119.0 |
| doi.org | 23.8 | 23.8 | 14.3 | 4.8 | 9.5 | 23.8 | 0.0 | 0.0 | 109.7 |
| mozilla.com | 10.4 | 6.2 | 35.4 | 8.3 | 10.4 | 6.2 | 16.7 | 6.2 | 98.0 |
| apache.org | 5.9 | 23.5 | 29.4 | 5.9 | 5.9 | 11.8 | 11.8 | 5.9 | 83.4 |
| rlcdn.com | 21.4 | 17.9 | 21.4 | 0.0 | 3.6 | 17.9 | 14.3 | 3.6 | 76.5 |
| doubleverify.com | 25.2 | 23.6 | 10.2 | 1.6 | 7.9 | 15.0 | 11.8 | 4.7 | 71.1 |
| adsafeprotected.com | 11.1 | 14.3 | 25.4 | 4.8 | 9.5 | 19.0 | 11.1 | 4.8 | 49.3 |
| bbc.co.uk | 10.0 | 16.7 | 23.3 | 3.3 | 6.7 | 20.0 | 10.0 | 10.0 | 46.8 |
| who.int | 8.8 | 17.6 | 23.5 | 10.3 | 4.4 | 17.6 | 11.8 | 5.9 | 43.3 |
| tds.net | 9.3 | 18.5 | 22.2 | 1.9 | 7.4 | 13.0 | 11.1 | 16.7 | 42.9 |
| mit.edu | 16.2 | 14.7 | 25.0 | 5.9 | 10.3 | 14.7 | 5.9 | 7.4 | 42.6 |
| opera.com | 13.1 | 17.9 | 22.0 | 5.4 | 13.7 | 16.7 | 7.7 | 3.6 | 41.7 |
| nginx.com | 23.4 | 13.2 | 20.4 | 6.4 | 11.9 | 11.1 | 7.2 | 6.4 | 41.0 |
| office.net | 14.9 | 19.9 | 19.9 | 2.5 | 8.7 | 16.8 | 10.6 | 6.8 | 40.5 |
| att.net | 16.7 | 16.7 | 14.8 | 1.9 | 9.3 | 22.2 | 9.3 | 9.3 | 39.9 |
| rubiconproject.com | 7.9 | 18.4 | 15.8 | 5.3 | 10.5 | 23.7 | 7.9 | 10.5 | 39.1 |
| force.com | 12.1 | 18.2 | 21.2 | 3.0 | 15.2 | 9.1 | 6.1 | 15.2 | 37.9 |
| ibm.com | 14.2 | 17.2 | 22.4 | 3.7 | 9.7 | 16.4 | 8.2 | 8.2 | 37.0 |
| bbc.com | 9.7 | 16.1 | 22.6 | 6.5 | 6.5 | 19.4 | 9.7 | 9.7 | 37.0 |
| nih.gov | 13.3 | 15.7 | 22.3 | 8.7 | 6.0 | 18.7 | 9.3 | 6.0 | 36.4 |
| launchdarkly.com | 13.9 | 19.8 | 18.8 | 7.9 | 5.9 | 17.8 | 10.9 | 5.0 | 35.3 |
| europa.eu | 16.7 | 8.3 | 20.8 | 16.7 | 8.3 | 8.3 | 16.7 | 4.2 | 34.7 |
| criteo.com | 17.1 | 20.7 | 15.9 | 4.9 | 11.0 | 15.9 | 8.5 | 6.1 | 32.6 |
| adsrvr.org | 11.3 | 15.1 | 20.8 | 3.8 | 13.2 | 18.9 | 7.5 | 9.4 | 32.5 |
| android.com | 12.0 | 17.1 | 18.8 | 4.3 | 16.2 | 16.2 | 11.1 | 4.3 | 32.3 |
| webex.com | 17.6 | 15.5 | 20.9 | 4.3 | 11.8 | 13.9 | 8.0 | 8.0 | 30.8 |
| ubuntu.com | 19.8 | 15.4 | 19.1 | 6.2 | 10.5 | 14.8 | 7.4 | 6.8 | 30.5 |
| pinterest.com | 14.1 | 18.6 | 18.8 | 3.9 | 7.2 | 16.6 | 11.9 | 8.9 | 30.4 |
| cisco.com | 17.7 | 17.2 | 19.8 | 5.2 | 10.9 | 14.1 | 8.3 | 6.8 | 30.1 |
| salesforce.com | 12.1 | 12.1 | 21.2 | 6.1 | 18.2 | 9.1 | 6.1 | 15.2 | 30.0 |
| msn.com | 17.1 | 20.2 | 16.5 | 5.7 | 7.0 | 15.5 | 9.0 | 9.1 | 29.5 |
| windows.net | 16.6 | 19.8 | 16.7 | 5.7 | 7.1 | 16.2 | 9.3 | 8.6 | 29.0 |
| www.gov.uk | 13.0 | 15.6 | 20.8 | 7.8 | 5.2 | 18.2 | 9.1 | 10.4 | 28.9 |
| outlook.com | 17.3 | 19.9 | 16.4 | 5.6 | 6.9 | 15.5 | 9.2 | 9.3 | 28.8 |
| office365.com | 17.3 | 19.9 | 16.5 | 5.6 | 7.3 | 15.4 | 9.0 | 9.0 | 28.8 |
| skype.com | 17.0 | 19.8 | 16.7 | 5.7 | 7.2 | 15.7 | 9.0 | 8.9 | 28.5 |
| snapchat.com | 5.8 | 18.5 | 16.0 | 6.2 | 14.5 | 18.5 | 12.7 | 7.6 | 28.3 |
| bit.ly | 11.1 | 19.2 | 18.8 | 5.4 | 11.5 | 16.9 | 10.7 | 6.5 | 27.9 |
| live.com | 17.0 | 19.7 | 16.6 | 5.7 | 7.4 | 15.6 | 9.0 | 9.1 | 27.9 |
| casalemedia.com | 12.9 | 18.6 | 15.7 | 1.4 | 11.4 | 17.1 | 11.4 | 11.4 | 27.8 |
| windows.com | 16.7 | 19.5 | 16.8 | 5.9 | 7.1 | 15.9 | 9.3 | 8.8 | 27.8 |
| grammarly.com | 16.8 | 17.7 | 18.2 | 4.1 | 12.3 | 14.5 | 9.5 | 6.8 | 27.7 |
| w3.org | 14.3 | 17.5 | 19.0 | 14.3 | 7.9 | 14.3 | 9.5 | 3.2 | 27.7 |
| office.com | 17.0 | 19.6 | 16.4 | 5.6 | 7.4 | 15.7 | 9.5 | 8.9 | 27.5 |
| sharepoint.com | 17.0 | 19.7 | 16.4 | 5.7 | 7.2 | 15.6 | 9.2 | 9.2 | 27.5 |
| harvard.edu | 16.4 | 17.9 | 13.4 | 3.0 | 11.9 | 17.9 | 7.5 | 11.9 | 27.3 |
| name-services.com | 17.2 | 13.8 | 6.9 | 6.9 | 13.8 | 20.7 | 13.8 | 6.9 | 27.0 |
| google.com | 13.6 | 19.9 | 18.2 | 5.5 | 11.4 | 15.3 | 7.2 | 8.9 | 26.8 |
| microsoftonline.com | 16.5 | 19.0 | 16.9 | 6.2 | 7.5 | 16.4 | 8.8 | 8.7 | 26.7 |
| sentry.io | 15.8 | 15.2 | 18.0 | 4.3 | 11.5 | 17.7 | 10.9 | 6.5 | 26.0 |
| hubspot.com | 13.2 | 17.1 | 20.6 | 5.7 | 9.2 | 15.8 | 10.5 | 7.9 | 25.8 |
| unity3d.com | 17.4 | 18.5 | 14.7 | 4.9 | 13.7 | 15.6 | 8.0 | 7.3 | 25.7 |
| kaspersky.com | 13.9 | 17.6 | 20.1 | 6.6 | 8.9 | 15.8 | 9.8 | 7.3 | 25.5 |
| dnsmadeeasy.com | 11.1 | 18.1 | 15.3 | 2.8 | 11.1 | 18.1 | 9.7 | 13.9 | 25.4 |
| wal-mart.com | 13.8 | 16.7 | 17.7 | 3.2 | 9.2 | 17.0 | 13.1 | 9.2 | 25.1 |
| zemanta.com | 19.0 | 18.0 | 15.0 | 8.0 | 10.0 | 15.0 | 10.0 | 5.0 | 24.9 |
| googlevideo.com | 14.2 | 19.7 | 17.2 | 5.4 | 11.3 | 15.5 | 7.1 | 9.6 | 24.8 |
| goo.gl | 14.3 | 20.0 | 18.0 | 6.5 | 10.6 | 14.3 | 6.9 | 9.4 | 24.6 |
| shopify.com | 15.3 | 16.5 | 16.5 | 2.7 | 9.2 | 17.2 | 11.6 | 11.1 | 24.5 |
| cnn.com | 17.2 | 13.1 | 14.2 | 4.1 | 18.7 | 13.9 | 12.0 | 6.7 | 24.3 |
| sharethrough.com | 13.7 | 18.8 | 16.4 | 4.7 | 14.6 | 15.3 | 10.0 | 6.6 | 24.3 |
| flickr.com | 11.1 | 20.7 | 17.3 | 5.8 | 11.5 | 15.4 | 9.1 | 9.1 | 24.2 |
| azurewebsites.net | 15.3 | 18.2 | 12.5 | 4.0 | 14.2 | 17.6 | 10.2 | 8.0 | 24.0 |
| youtu.be | 15.0 | 19.4 | 17.8 | 6.5 | 10.9 | 14.2 | 7.3 | 8.9 | 23.4 |
| digicert.com | 15.8 | 16.7 | 16.3 | 3.8 | 7.2 | 15.8 | 13.9 | 10.5 | 23.3 |
| zoom.us | 18.3 | 18.7 | 14.3 | 5.1 | 9.2 | 15.0 | 9.3 | 10.1 | 23.3 |
| sourceforge.net | 16.3 | 14.4 | 19.9 | 5.8 | 14.4 | 12.3 | 10.2 | 6.6 | 23.3 |
| mcafee.com | 12.7 | 18.7 | 13.3 | 3.0 | 12.7 | 17.5 | 12.7 | 9.6 | 23.1 |
| ebay.com | 10.5 | 16.3 | 22.1 | 11.6 | 7.0 | 12.8 | 8.1 | 11.6 | 23.1 |
| facebook.com | 10.6 | 19.5 | 12.9 | 3.5 | 9.8 | 16.2 | 14.6 | 12.9 | 22.8 |
| smartadserver.com | 18.3 | 18.3 | 12.7 | 6.3 | 11.1 | 16.7 | 8.7 | 7.9 | 22.6 |
| wikimedia.org | 17.3 | 16.8 | 18.5 | 7.0 | 8.8 | 14.0 | 10.5 | 7.0 | 22.6 |
| reddit.com | 7.1 | 17.2 | 19.2 | 6.1 | 13.6 | 15.7 | 9.6 | 11.6 | 22.6 |
| roblox.com | 15.3 | 16.7 | 18.1 | 4.0 | 13.9 | 14.1 | 8.7 | 9.3 | 22.5 |
| mzstatic.com | 14.6 | 19.2 | 13.6 | 4.9 | 14.6 | 16.0 | 9.1 | 8.0 | 22.5 |
| googletagmanager.com | 13.9 | 19.3 | 18.0 | 6.6 | 11.5 | 14.3 | 7.4 | 9.0 | 22.4 |
| google.co.uk | 13.6 | 19.8 | 18.1 | 7.0 | 10.7 | 14.0 | 7.4 | 9.5 | 22.3 |
| scorecardresearch.com | 14.5 | 15.3 | 9.9 | 6.9 | 16.8 | 19.8 | 8.4 | 8.4 | 22.3 |
| youtube.com | 14.0 | 19.3 | 17.7 | 6.2 | 11.5 | 14.4 | 7.4 | 9.5 | 22.3 |

**Table 8: Website privacy policies ranked by the variance of the percentages of individual parameter types out of all annotations *(continued on next page).***

| Website | % Aim | % Attribute | % Condition | % Consequence | % Modality | % Recipient | % Sender | % Subject | Variance |
|---|---|---|---|---|---|---|---|---|---|
| pubmatic.com | 16.08 | 15.50 | 14.04 | 3.51 | 16.67 | 15.79 | 9.65 | 8.77 | 22.17 |
| ampproject.org | 18.23 | 18.42 | 14.36 | 5.34 | 8.84 | 14.92 | 9.58 | 10.31 | 22.17 |
| googleapis.com | 14.34 | 18.85 | 18.44 | 6.97 | 10.66 | 14.34 | 7.79 | 8.61 | 21.98 |
| bing.com | 16.67 | 17.65 | 15.20 | 4.90 | 9.31 | 16.67 | 10.78 | 8.82 | 21.83 |
| 3lift.com | 10.64 | 16.60 | 18.30 | 6.38 | 14.47 | 16.17 | 11.49 | 5.96 | 21.78 |
| b-cdn.net | 17.81 | 16.44 | 15.07 | 2.74 | 12.33 | 13.70 | 10.96 | 10.96 | 21.68 |
| applovin.com | 12.57 | 17.28 | 16.23 | 4.71 | 16.23 | 15.71 | 8.90 | 8.38 | 21.65 |
| azure.com | 14.36 | 17.68 | 14.36 | 3.87 | 13.81 | 17.13 | 9.94 | 8.84 | 21.62 |
| xiaomi.com | 17.92 | 17.38 | 17.20 | 5.73 | 10.75 | 13.08 | 9.68 | 8.24 | 21.52 |
| wikipedia.org | 16.70 | 16.13 | 18.98 | 7.78 | 8.16 | 14.61 | 10.44 | 7.21 | 21.51 |
| fastly.net | 16.89 | 17.43 | 18.23 | 5.36 | 9.38 | 12.87 | 9.92 | 9.92 | 21.50 |
| bidswitch.net | 10.58 | 12.50 | 19.23 | 3.85 | 17.31 | 13.46 | 11.54 | 11.54 | 21.40 |
| shipt.com | 10.78 | 17.65 | 17.65 | 5.39 | 11.27 | 16.18 | 13.73 | 7.35 | 21.35 |
| amazon.com | 11.67 | 17.22 | 20.56 | 8.33 | 9.44 | 15.00 | 10.00 | 7.78 | 21.34 |
| googleadservices.com | 13.68 | 19.23 | 17.52 | 6.41 | 10.68 | 14.96 | 7.69 | 9.83 | 21.22 |
| gstatic.com | 12.45 | 18.67 | 18.26 | 7.05 | 10.79 | 15.77 | 7.47 | 9.54 | 21.22 |
| reuters.com | 15.79 | 18.42 | 16.27 | 5.74 | 12.68 | 14.59 | 9.09 | 7.42 | 21.09 |
| netflix.com | 16.52 | 17.86 | 16.07 | 5.36 | 11.61 | 15.18 | 9.38 | 8.04 | 20.90 |
| sciencedirect.com | 12.77 | 16.67 | 20.57 | 5.67 | 10.64 | 12.77 | 8.87 | 12.06 | 20.87 |
| github.io | 11.41 | 17.72 | 18.53 | 5.70 | 11.61 | 16.09 | 8.35 | 10.59 | 20.80 |
| booking.com | 15.41 | 18.24 | 15.25 | 4.25 | 11.64 | 15.72 | 10.38 | 9.12 | 20.60 |
| doubleclick.net | 13.62 | 18.72 | 17.45 | 5.96 | 11.06 | 15.32 | 8.09 | 9.79 | 20.59 |
| github.com | 11.48 | 17.83 | 18.03 | 5.53 | 11.07 | 16.60 | 9.02 | 10.45 | 20.54 |
| ui.com | 13.64 | 18.56 | 12.50 | 6.06 | 15.15 | 16.67 | 10.98 | 6.44 | 20.42 |
| theguardian.com | 16.75 | 16.75 | 16.75 | 7.39 | 13.79 | 14.29 | 6.90 | 7.39 | 20.38 |
| google-analytics.com | 14.80 | 18.80 | 17.60 | 6.80 | 11.20 | 14.00 | 7.60 | 9.20 | 20.33 |
| hotjar.com | 13.01 | 17.07 | 17.07 | 3.25 | 10.57 | 15.45 | 12.20 | 11.38 | 20.20 |
| yahoo.com | 10.77 | 15.38 | 16.92 | 3.08 | 15.38 | 14.62 | 10.00 | 13.85 | 20.08 |
| comcast.net | 15.03 | 17.34 | 15.03 | 5.20 | 11.56 | 16.76 | 12.14 | 6.94 | 19.95 |
| appsflyer.com | 16.27 | 13.25 | 18.67 | 6.02 | 10.24 | 15.66 | 12.65 | 7.23 | 19.67 |
| gandi.net | 12.06 | 15.00 | 18.24 | 7.35 | 16.76 | 15.00 | 8.53 | 7.06 | 19.38 |
| t.co | 14.94 | 18.31 | 16.49 | 5.84 | 8.70 | 15.45 | 9.61 | 10.65 | 19.27 |
| apple.com | 15.09 | 17.54 | 14.39 | 5.26 | 14.39 | 15.79 | 9.47 | 8.07 | 18.75 |
| wordpress.org | 12.75 | 17.45 | 19.46 | 8.05 | 10.74 | 14.09 | 10.07 | 7.38 | 18.65 |
| registrar-servers.com | 14.81 | 15.56 | 17.04 | 6.67 | 6.67 | 17.04 | 11.85 | 10.37 | 18.41 |
| hp.com | 11.43 | 15.92 | 14.69 | 4.90 | 15.51 | 17.14 | 12.65 | 7.76 | 18.37 |
| tiktok.com | 10.96 | 18.26 | 14.61 | 3.65 | 11.87 | 15.53 | 12.79 | 12.33 | 18.28 |
| frontapp.com | 16.30 | 14.13 | 17.93 | 8.15 | 10.33 | 16.85 | 7.61 | 8.70 | 18.23 |
| amazonaws.com | 11.11 | 16.16 | 17.68 | 5.56 | 11.11 | 17.17 | 9.09 | 12.12 | 17.98 |
| tinyurl.com | 15.38 | 15.38 | 13.46 | 5.77 | 15.38 | 15.38 | 13.46 | 5.77 | 17.96 |
| netflix.net | 15.62 | 16.52 | 16.52 | 6.25 | 12.05 | 15.62 | 10.27 | 7.14 | 17.82 |
| go.com | 8.84 | 17.01 | 17.35 | 5.78 | 11.90 | 16.33 | 12.93 | 9.86 | 17.78 |
| dailymail.co.uk | 15.64 | 17.82 | 14.55 | 5.27 | 9.82 | 16.18 | 10.36 | 10.36 | 17.77 |
| intuit.com | 15.03 | 15.03 | 16.26 | 3.68 | 13.80 | 15.34 | 10.12 | 10.74 | 17.60 |
| creativecommons.org | 10.04 | 16.87 | 16.87 | 5.22 | 15.66 | 14.46 | 11.65 | 9.24 | 17.49 |
| amazon.co.uk | 15.35 | 15.35 | 17.21 | 6.51 | 9.77 | 16.28 | 7.44 | 12.09 | 17.40 |
| icloud.com | 15.02 | 18.03 | 13.73 | 6.01 | 13.30 | 15.88 | 10.30 | 7.73 | 17.18 |
| vimeo.com | 8.46 | 16.05 | 16.70 | 5.42 | 14.32 | 16.27 | 10.20 | 12.58 | 17.07 |
| spotify.com | 14.41 | 17.94 | 17.06 | 5.00 | 10.88 | 12.06 | 10.00 | 12.65 | 17.05 |
| soundcloud.com | 11.15 | 16.07 | 20.33 | 7.54 | 9.84 | 13.77 | 9.84 | 11.48 | 16.75 |
| taboola.com | 13.53 | 17.16 | 11.55 | 4.29 | 14.19 | 16.50 | 12.87 | 9.90 | 16.69 |
| epicgames.com | 9.09 | 15.34 | 19.60 | 7.10 | 15.06 | 12.50 | 11.36 | 9.94 | 16.23 |
| researchgate.net | 16.13 | 17.42 | 16.13 | 8.06 | 9.35 | 14.19 | 11.61 | 7.10 | 16.14 |
| nytimes.com | 13.43 | 16.72 | 17.31 | 6.57 | 7.76 | 15.52 | 11.94 | 10.75 | 16.00 |
| weebly.com | 16.21 | 15.17 | 18.28 | 7.59 | 11.72 | 13.79 | 9.31 | 7.93 | 15.96 |
| aol.com | 7.69 | 15.38 | 15.38 | 11.54 | 11.54 | 19.23 | 11.54 | 7.69 | 15.85 |
| twitch.tv | 10.34 | 15.52 | 16.09 | 5.75 | 14.94 | 16.67 | 12.07 | 8.62 | 15.83 |
| macromedia.com | 12.74 | 16.88 | 18.15 | 7.32 | 12.74 | 14.65 | 8.92 | 8.60 | 15.81 |
| imdb.com | 12.50 | 14.13 | 16.85 | 7.07 | 10.33 | 17.93 | 13.59 | 7.61 | 15.78 |
| issuu.com | 14.81 | 16.40 | 15.87 | 4.76 | 13.23 | 14.81 | 9.52 | 10.58 | 15.75 |
| adobe.com | 12.06 | 17.14 | 18.41 | 7.62 | 12.70 | 14.29 | 9.21 | 8.57 | 15.72 |
| washingtonpost.com | 13.53 | 15.84 | 14.19 | 5.28 | 17.16 | 14.19 | 10.56 | 9.24 | 15.17 |
| forbes.com | 14.08 | 14.93 | 16.90 | 6.48 | 9.86 | 16.34 | 13.24 | 8.17 | 15.01 |
| deviantart.com | 10.68 | 16.83 | 17.80 | 7.77 | 12.30 | 15.86 | 9.39 | 9.39 | 14.75 |
| paypal.com | 12.75 | 16.17 | 16.06 | 4.37 | 13.11 | 14.76 | 11.81 | 10.98 | 14.37 |
| linkedin.com | 13.09 | 15.88 | 16.99 | 5.85 | 10.31 | 16.16 | 10.86 | 10.86 | 14.24 |
| cdc.gov | 13.25 | 16.06 | 18.47 | 6.83 | 12.05 | 13.25 | 11.65 | 8.43 | 14.17 |
| badoo.com | 17.24 | 14.73 | 16.61 | 9.72 | 9.40 | 15.05 | 9.09 | 8.15 | 14.11 |
| slideshare.net | 12.91 | 16.21 | 16.76 | 5.49 | 10.44 | 15.66 | 11.54 | 10.99 | 14.08 |
| wp.com | 12.36 | 16.22 | 16.60 | 5.41 | 10.81 | 15.83 | 11.20 | 11.58 | 13.93 |
| samsung.com | 13.85 | 15.82 | 12.53 | 5.49 | 14.73 | 16.70 | 10.11 | 10.77 | 13.32 |
| instagram.com | 12.99 | 16.14 | 15.75 | 7.09 | 14.57 | 14.96 | 11.02 | 7.48 | 13.01 |
| wordpress.com | 13.14 | 15.69 | 16.27 | 5.69 | 10.98 | 15.88 | 10.98 | 11.37 | 12.65 |
| gravatar.com | 12.77 | 15.91 | 16.50 | 5.70 | 11.00 | 15.52 | 11.39 | 11.20 | 12.62 |
| espn.com | 8.63 | 16.31 | 15.59 | 8.63 | 15.35 | 14.63 | 12.71 | 8.15 | 12.22 |
| tumblr.com | 12.21 | 16.54 | 16.79 | 8.40 | 12.72 | 15.27 | 9.41 | 8.65 | 11.96 |
| dropbox.com | 13.27 | 15.49 | 15.49 | 7.96 | 13.27 | 16.37 | 10.18 | 7.96 | 11.51 |
| cloudflare.com | 11.45 | 17.18 | 17.62 | 7.93 | 11.89 | 12.78 | 10.13 | 11.01 | 11.20 |
| medium.com | 10.92 | 15.28 | 16.59 | 9.17 | 11.79 | 16.59 | 8.30 | 11.35 | 10.62 |
| newrelic.com | 12.16 | 14.86 | 18.02 | 6.76 | 11.71 | 12.61 | 11.26 | 12.61 | 10.19 |
| hipages.com.au | 12.28 | 12.28 | 12.28 | 7.02 | 14.04 | 17.54 | 12.28 | 12.28 | 8.30 |

**Table 9: Website privacy policies ranked by the variance of the percentages of individual parameter types out of all annotations *(continued from previous page).***

# K Parameter Density Data

| Website | $\frac{\#\ Aims}{\#\ Sentences}$ | $\frac{\#\ Attributes}{\#\ Sentences}$ | $\frac{\#\ Conditions}{\#\ Sentences}$ | $\frac{\#\ Consequences}{\#\ Sentences}$ | $\frac{\#\ Modalities}{\#\ Sentences}$ | $\frac{\#\ Recipients}{\#\ Sentences}$ | $\frac{\#\ Senders}{\#\ Sentences}$ | $\frac{\#\ Subjects}{\#\ Sentences}$ | **Total # parameters / # Sentences** |
|---|---|---|---|---|---|---|---|---|---|
| b-cdn.net | 0.0774 | 0.0714 | 0.0655 | 0.0119 | 0.0536 | 0.0595 | 0.0476 | 0.0476 | 0.4345 |
| facebook.com | 0.0389 | 0.0716 | 0.0473 | 0.0130 | 0.0358 | 0.0595 | 0.0534 | 0.0473 | 0.3666 |
| bing.com | 0.0599 | 0.0634 | 0.0546 | 0.0176 | 0.0335 | 0.0599 | 0.0387 | 0.0317 | 0.3592 |
| slideshare.net | 0.0459 | 0.0576 | 0.0596 | 0.0195 | 0.0371 | 0.0557 | 0.0410 | 0.0391 | 0.3555 |
| linkedin.com | 0.0459 | 0.0557 | 0.0596 | 0.0205 | 0.0361 | 0.0566 | 0.0381 | 0.0381 | 0.3506 |
| zoom.us | 0.0641 | 0.0654 | 0.0500 | 0.0179 | 0.0321 | 0.0526 | 0.0327 | 0.0353 | 0.3500 |
| ampproject.org | 0.0635 | 0.0641 | 0.0500 | 0.0186 | 0.0308 | 0.0519 | 0.0333 | 0.0359 | 0.3481 |
| sciencedirect.com | 0.0437 | 0.0570 | 0.0704 | 0.0194 | 0.0364 | 0.0437 | 0.0303 | 0.0413 | 0.3422 |
| snapchat.com | 0.0183 | 0.0585 | 0.0505 | 0.0195 | 0.0459 | 0.0585 | 0.0401 | 0.0241 | 0.3154 |
| t.co | 0.0467 | 0.0572 | 0.0515 | 0.0183 | 0.0272 | 0.0483 | 0.0300 | 0.0333 | 0.3125 |
| tumblr.com | 0.0375 | 0.0508 | 0.0516 | 0.0258 | 0.0391 | 0.0469 | 0.0289 | 0.0266 | 0.3070 |
| icloud.com | 0.0437 | 0.0525 | 0.0400 | 0.0175 | 0.0387 | 0.0462 | 0.0300 | 0.0225 | 0.2913 |
| spotify.com | 0.0417 | 0.0519 | 0.0493 | 0.0145 | 0.0315 | 0.0349 | 0.0289 | 0.0366 | 0.2891 |
| issuu.com | 0.0427 | 0.0473 | 0.0457 | 0.0137 | 0.0381 | 0.0427 | 0.0274 | 0.0305 | 0.2881 |
| samsung.com | 0.0398 | 0.0455 | 0.0360 | 0.0158 | 0.0423 | 0.0480 | 0.0290 | 0.0309 | 0.2872 |
| mzstatic.com | 0.0420 | 0.0550 | 0.0390 | 0.0140 | 0.0420 | 0.0460 | 0.0260 | 0.0230 | 0.2870 |
| creativecommons.org | 0.0287 | 0.0482 | 0.0482 | 0.0149 | 0.0447 | 0.0413 | 0.0333 | 0.0264 | 0.2856 |
| apple.com | 0.0430 | 0.0500 | 0.0410 | 0.0150 | 0.0410 | 0.0450 | 0.0270 | 0.0230 | 0.2850 |
| forbes.com | 0.0398 | 0.0422 | 0.0478 | 0.0183 | 0.0279 | 0.0462 | 0.0374 | 0.0231 | 0.2826 |
| azure.com | 0.0401 | 0.0494 | 0.0401 | 0.0108 | 0.0386 | 0.0478 | 0.0278 | 0.0247 | 0.2793 |
| reddit.com | 0.0197 | 0.0478 | 0.0534 | 0.0169 | 0.0379 | 0.0435 | 0.0267 | 0.0323 | 0.2781 |
| wp.com | 0.0339 | 0.0445 | 0.0456 | 0.0148 | 0.0297 | 0.0434 | 0.0307 | 0.0318 | 0.2744 |
| azurewebsites.net | 0.0417 | 0.0494 | 0.0340 | 0.0108 | 0.0386 | 0.0478 | 0.0278 | 0.0216 | 0.2716 |
| wordpress.com | 0.0355 | 0.0424 | 0.0440 | 0.0154 | 0.0297 | 0.0429 | 0.0297 | 0.0307 | 0.2701 |
| epicgames.com | 0.0245 | 0.0414 | 0.0529 | 0.0192 | 0.0406 | 0.0337 | 0.0307 | 0.0268 | 0.2699 |
| gravatar.com | 0.0344 | 0.0429 | 0.0445 | 0.0154 | 0.0297 | 0.0418 | 0.0307 | 0.0302 | 0.2696 |
| vimeo.com | 0.0228 | 0.0432 | 0.0450 | 0.0146 | 0.0386 | 0.0438 | 0.0275 | 0.0339 | 0.2693 |
| instagram.com | 0.0350 | 0.0434 | 0.0424 | 0.0191 | 0.0392 | 0.0403 | 0.0297 | 0.0201 | 0.2691 |
| shopify.com | 0.0410 | 0.0443 | 0.0443 | 0.0072 | 0.0247 | 0.0462 | 0.0312 | 0.0299 | 0.2689 |
| goo.gl | 0.0377 | 0.0528 | 0.0474 | 0.0172 | 0.0280 | 0.0377 | 0.0183 | 0.0248 | 0.2640 |
| google-analytics.com | 0.0389 | 0.0494 | 0.0462 | 0.0179 | 0.0294 | 0.0368 | 0.0200 | 0.0242 | 0.2626 |
| ebay.com | 0.0274 | 0.0427 | 0.0579 | 0.0305 | 0.0183 | 0.0335 | 0.0213 | 0.0305 | 0.2622 |
| adobe.com | 0.0315 | 0.0447 | 0.0480 | 0.0199 | 0.0331 | 0.0373 | 0.0240 | 0.0224 | 0.2608 |
| tiktok.com | 0.0286 | 0.0476 | 0.0381 | 0.0095 | 0.0310 | 0.0405 | 0.0333 | 0.0321 | 0.2607 |
| comcast.net | 0.0392 | 0.0452 | 0.0392 | 0.0136 | 0.0301 | 0.0437 | 0.0316 | 0.0181 | 0.2605 |
| macromedia.com | 0.0331 | 0.0439 | 0.0472 | 0.0190 | 0.0331 | 0.0381 | 0.0232 | 0.0224 | 0.2599 |
| google.co.uk | 0.0353 | 0.0513 | 0.0470 | 0.0182 | 0.0278 | 0.0363 | 0.0192 | 0.0246 | 0.2596 |
| youtu.be | 0.0389 | 0.0504 | 0.0462 | 0.0168 | 0.0284 | 0.0368 | 0.0189 | 0.0231 | 0.2595 |
| googlevideo.com | 0.0366 | 0.0506 | 0.0442 | 0.0140 | 0.0291 | 0.0399 | 0.0183 | 0.0248 | 0.2575 |
| gstatic.com | 0.0321 | 0.0481 | 0.0470 | 0.0182 | 0.0278 | 0.0406 | 0.0192 | 0.0246 | 0.2575 |
| flickr.com | 0.0285 | 0.0532 | 0.0446 | 0.0149 | 0.0297 | 0.0396 | 0.0235 | 0.0235 | 0.2574 |
| cdc.gov | 0.0341 | 0.0413 | 0.0475 | 0.0176 | 0.0310 | 0.0341 | 0.0300 | 0.0217 | 0.2572 |
| dropbox.com | 0.0341 | 0.0398 | 0.0398 | 0.0205 | 0.0341 | 0.0420 | 0.0261 | 0.0205 | 0.2568 |
| googleapis.com | 0.0368 | 0.0483 | 0.0473 | 0.0179 | 0.0273 | 0.0368 | 0.0200 | 0.0221 | 0.2563 |
| googletagmanager.com | 0.0357 | 0.0494 | 0.0462 | 0.0168 | 0.0294 | 0.0368 | 0.0189 | 0.0231 | 0.2563 |
| doubleverify.com | 0.0645 | 0.0605 | 0.0262 | 0.0040 | 0.0202 | 0.0383 | 0.0302 | 0.0121 | 0.2560 |
| youtube.com | 0.0357 | 0.0494 | 0.0452 | 0.0158 | 0.0294 | 0.0368 | 0.0189 | 0.0242 | 0.2553 |
| google.com | 0.0345 | 0.0506 | 0.0463 | 0.0140 | 0.0291 | 0.0388 | 0.0183 | 0.0226 | 0.2543 |
| microsoftonline.com | 0.0415 | 0.0479 | 0.0426 | 0.0156 | 0.0189 | 0.0412 | 0.0221 | 0.0218 | 0.2516 |
| doubleclick.net | 0.0342 | 0.0470 | 0.0438 | 0.0150 | 0.0278 | 0.0385 | 0.0203 | 0.0246 | 0.2511 |
| windows.net | 0.0415 | 0.0498 | 0.0420 | 0.0143 | 0.0177 | 0.0406 | 0.0234 | 0.0215 | 0.2508 |
| github.io | 0.0286 | 0.0444 | 0.0464 | 0.0143 | 0.0291 | 0.0403 | 0.0209 | 0.0265 | 0.2505 |
| googleadservices.com | 0.0342 | 0.0481 | 0.0438 | 0.0160 | 0.0267 | 0.0374 | 0.0192 | 0.0246 | 0.2500 |
| windows.com | 0.0417 | 0.0487 | 0.0420 | 0.0147 | 0.0177 | 0.0395 | 0.0231 | 0.0219 | 0.2494 |
| dailymail.co.uk | 0.0389 | 0.0444 | 0.0362 | 0.0131 | 0.0245 | 0.0403 | 0.0258 | 0.0258 | 0.2491 |
| github.com | 0.0286 | 0.0444 | 0.0449 | 0.0138 | 0.0276 | 0.0413 | 0.0224 | 0.0260 | 0.2490 |
| wordpress.org | 0.0317 | 0.0433 | 0.0483 | 0.0200 | 0.0267 | 0.0350 | 0.0250 | 0.0183 | 0.2483 |
| imdb.com | 0.0309 | 0.0349 | 0.0417 | 0.0175 | 0.0255 | 0.0444 | 0.0336 | 0.0188 | 0.2473 |
| bit.ly | 0.0275 | 0.0473 | 0.0464 | 0.0133 | 0.0284 | 0.0417 | 0.0265 | 0.0161 | 0.2472 |
| pinterest.com | 0.0346 | 0.0455 | 0.0462 | 0.0095 | 0.0177 | 0.0408 | 0.0292 | 0.0217 | 0.2452 |
| medium.com | 0.0265 | 0.0371 | 0.0403 | 0.0222 | 0.0286 | 0.0403 | 0.0201 | 0.0275 | 0.2426 |
| soundcloud.com | 0.0269 | 0.0388 | 0.0491 | 0.0182 | 0.0237 | 0.0332 | 0.0237 | 0.0277 | 0.2413 |
| live.com | 0.0409 | 0.0474 | 0.0399 | 0.0136 | 0.0177 | 0.0375 | 0.0216 | 0.0220 | 0.2407 |
| office.com | 0.0410 | 0.0471 | 0.0393 | 0.0134 | 0.0177 | 0.0378 | 0.0228 | 0.0214 | 0.2406 |
| sharepoint.com | 0.0409 | 0.0473 | 0.0394 | 0.0138 | 0.0173 | 0.0375 | 0.0221 | 0.0222 | 0.2406 |
| office365.com | 0.0415 | 0.0479 | 0.0397 | 0.0135 | 0.0176 | 0.0369 | 0.0215 | 0.0216 | 0.2401 |
| deviantart.com | 0.0256 | 0.0404 | 0.0427 | 0.0186 | 0.0295 | 0.0380 | 0.0225 | 0.0225 | 0.2399 |
| harvard.edu | 0.0393 | 0.0429 | 0.0321 | 0.0071 | 0.0286 | 0.0429 | 0.0179 | 0.0286 | 0.2393 |
| outlook.com | 0.0414 | 0.0475 | 0.0393 | 0.0133 | 0.0165 | 0.0370 | 0.0221 | 0.0222 | 0.2392 |
| skype.com | 0.0405 | 0.0472 | 0.0398 | 0.0137 | 0.0172 | 0.0376 | 0.0215 | 0.0213 | 0.2388 |
| msn.com | 0.0409 | 0.0481 | 0.0393 | 0.0135 | 0.0166 | 0.0370 | 0.0215 | 0.0217 | 0.2388 |
| applovin.com | 0.0297 | 0.0408 | 0.0384 | 0.0111 | 0.0384 | 0.0371 | 0.0210 | 0.0198 | 0.2364 |
| twitch.tv | 0.0242 | 0.0363 | 0.0376 | 0.0134 | 0.0349 | 0.0390 | 0.0282 | 0.0202 | 0.2339 |
| badoo.com | 0.0402 | 0.0344 | 0.0387 | 0.0227 | 0.0219 | 0.0351 | 0.0212 | 0.0190 | 0.2332 |
| yahoo.com | 0.0250 | 0.0357 | 0.0393 | 0.0071 | 0.0357 | 0.0339 | 0.0232 | 0.0321 | 0.2321 |
| android.com | 0.0278 | 0.0397 | 0.0437 | 0.0099 | 0.0377 | 0.0377 | 0.0258 | 0.0099 | 0.2321 |
| fastly.net | 0.0392 | 0.0404 | 0.0423 | 0.0124 | 0.0218 | 0.0299 | 0.0230 | 0.0230 | 0.2320 |
| booking.com | 0.0354 | 0.0419 | 0.0350 | 0.0098 | 0.0267 | 0.0361 | 0.0238 | 0.0210 | 0.2298 |
| xiaomi.com | 0.0411 | 0.0399 | 0.0395 | 0.0132 | 0.0247 | 0.0300 | 0.0222 | 0.0189 | 0.2294 |
| researchgate.net | 0.0370 | 0.0399 | 0.0370 | 0.0185 | 0.0214 | 0.0325 | 0.0266 | 0.0163 | 0.2293 |

**Table 10: Website privacy policies ranked by the total ratio of the number of annotated parameters to the number of sentences in the policy** ***(continued on next page).***

| Website | $\frac{\#\ Aims}{\#\ Sentences}$ | $\frac{\#\ Attributes}{\#\ Sentences}$ | $\frac{\#\ Conditions}{\#\ Sentences}$ | $\frac{\#\ Consequences}{\#\ Sentences}$ | $\frac{\#\ Modalities}{\#\ Sentences}$ | $\frac{\#\ Recipients}{\#\ Sentences}$ | $\frac{\#\ Senders}{\#\ Sentences}$ | $\frac{\#\ Subjects}{\#\ Sentences}$ | **Total # parameters / # Sentences** |
|---|---|---|---|---|---|---|---|---|---|
| paypal.com | 0.0291 | 0.0369 | 0.0366 | 0.0100 | 0.0299 | 0.0337 | 0.0269 | 0.0251 | 0.2282 |
| name-services.com | 0.0391 | 0.0312 | 0.0156 | 0.0156 | 0.0312 | 0.0469 | 0.0312 | 0.0156 | 0.2266 |
| dnsmadeeasy.com | 0.0250 | 0.0406 | 0.0344 | 0.0063 | 0.0250 | 0.0406 | 0.0219 | 0.0312 | 0.2250 |
| tds.net | 0.0208 | 0.0417 | 0.0500 | 0.0042 | 0.0167 | 0.0292 | 0.0250 | 0.0375 | 0.2250 |
| www.gov.uk | 0.0291 | 0.0349 | 0.0465 | 0.0174 | 0.0116 | 0.0407 | 0.0203 | 0.0233 | 0.2238 |
| ibm.com | 0.0317 | 0.0383 | 0.0500 | 0.0083 | 0.0217 | 0.0367 | 0.0183 | 0.0183 | 0.2233 |
| digicert.com | 0.0344 | 0.0365 | 0.0354 | 0.0083 | 0.0156 | 0.0344 | 0.0302 | 0.0229 | 0.2177 |
| weebly.com | 0.0352 | 0.0329 | 0.0397 | 0.0165 | 0.0254 | 0.0299 | 0.0202 | 0.0172 | 0.2171 |
| hotjar.com | 0.0282 | 0.0370 | 0.0370 | 0.0070 | 0.0229 | 0.0335 | 0.0264 | 0.0246 | 0.2165 |
| amazon.com | 0.0252 | 0.0373 | 0.0445 | 0.0180 | 0.0204 | 0.0325 | 0.0216 | 0.0168 | 0.2163 |
| registrar-servers.com | 0.0321 | 0.0337 | 0.0369 | 0.0144 | 0.0144 | 0.0369 | 0.0256 | 0.0224 | 0.2163 |
| scorecardresearch.com | 0.0312 | 0.0329 | 0.0214 | 0.0148 | 0.0362 | 0.0428 | 0.0181 | 0.0181 | 0.2155 |
| cnn.com | 0.0371 | 0.0282 | 0.0306 | 0.0089 | 0.0403 | 0.0298 | 0.0258 | 0.0145 | 0.2153 |
| roblox.com | 0.0328 | 0.0358 | 0.0388 | 0.0085 | 0.0299 | 0.0303 | 0.0188 | 0.0201 | 0.2150 |
| 3lift.com | 0.0228 | 0.0356 | 0.0392 | 0.0137 | 0.0310 | 0.0347 | 0.0246 | 0.0128 | 0.2144 |
| unity3d.com | 0.0371 | 0.0394 | 0.0313 | 0.0104 | 0.0293 | 0.0334 | 0.0172 | 0.0155 | 0.2136 |
| w3.org | 0.0304 | 0.0372 | 0.0405 | 0.0304 | 0.0169 | 0.0304 | 0.0203 | 0.0068 | 0.2128 |
| opera.com | 0.0278 | 0.0379 | 0.0467 | 0.0114 | 0.0290 | 0.0354 | 0.0164 | 0.0076 | 0.2121 |
| taboola.com | 0.0286 | 0.0363 | 0.0244 | 0.0091 | 0.0300 | 0.0349 | 0.0272 | 0.0209 | 0.2116 |
| wikipedia.org | 0.0350 | 0.0338 | 0.0398 | 0.0163 | 0.0171 | 0.0307 | 0.0219 | 0.0151 | 0.2098 |
| tinyurl.com | 0.0323 | 0.0323 | 0.0282 | 0.0121 | 0.0323 | 0.0323 | 0.0282 | 0.0121 | 0.2097 |
| pubmatic.com | 0.0337 | 0.0325 | 0.0294 | 0.0074 | 0.0349 | 0.0331 | 0.0202 | 0.0184 | 0.2096 |
| ui.com | 0.0285 | 0.0388 | 0.0261 | 0.0127 | 0.0316 | 0.0348 | 0.0229 | 0.0134 | 0.2089 |
| office.net | 0.0309 | 0.0412 | 0.0412 | 0.0052 | 0.0180 | 0.0348 | 0.0219 | 0.0142 | 0.2075 |
| adsafeprotected.com | 0.0230 | 0.0296 | 0.0526 | 0.0099 | 0.0197 | 0.0395 | 0.0230 | 0.0099 | 0.2072 |
| intuit.com | 0.0311 | 0.0311 | 0.0336 | 0.0076 | 0.0286 | 0.0317 | 0.0209 | 0.0222 | 0.2069 |
| amazonaws.com | 0.0229 | 0.0333 | 0.0365 | 0.0115 | 0.0229 | 0.0354 | 0.0187 | 0.0250 | 0.2062 |
| netflix.net | 0.0322 | 0.0340 | 0.0340 | 0.0129 | 0.0248 | 0.0322 | 0.0211 | 0.0147 | 0.2059 |
| netflix.com | 0.0340 | 0.0368 | 0.0331 | 0.0110 | 0.0239 | 0.0312 | 0.0193 | 0.0165 | 0.2059 |
| smartadserver.com | 0.0373 | 0.0373 | 0.0260 | 0.0130 | 0.0227 | 0.0341 | 0.0179 | 0.0162 | 0.2045 |
| gandi.net | 0.0246 | 0.0306 | 0.0373 | 0.0150 | 0.0343 | 0.0306 | 0.0174 | 0.0144 | 0.2043 |
| nih.gov | 0.0272 | 0.0319 | 0.0455 | 0.0177 | 0.0122 | 0.0380 | 0.0190 | 0.0122 | 0.2038 |
| washingtonpost.com | 0.0276 | 0.0323 | 0.0289 | 0.0108 | 0.0349 | 0.0289 | 0.0215 | 0.0188 | 0.2036 |
| wikimedia.org | 0.0353 | 0.0341 | 0.0377 | 0.0143 | 0.0179 | 0.0286 | 0.0214 | 0.0143 | 0.2036 |
| ubuntu.com | 0.0400 | 0.0312 | 0.0387 | 0.0125 | 0.0213 | 0.0300 | 0.0150 | 0.0138 | 0.2025 |
| go.com | 0.0179 | 0.0343 | 0.0350 | 0.0117 | 0.0240 | 0.0330 | 0.0261 | 0.0199 | 0.2019 |
| cloudflare.com | 0.0230 | 0.0346 | 0.0355 | 0.0160 | 0.0239 | 0.0257 | 0.0204 | 0.0222 | 0.2012 |
| hubspot.com | 0.0264 | 0.0343 | 0.0414 | 0.0114 | 0.0185 | 0.0317 | 0.0211 | 0.0158 | 0.2007 |
| kaspersky.com | 0.0279 | 0.0352 | 0.0402 | 0.0132 | 0.0178 | 0.0317 | 0.0197 | 0.0147 | 0.2005 |
| wal-mart.com | 0.0275 | 0.0332 | 0.0353 | 0.0064 | 0.0184 | 0.0339 | 0.0261 | 0.0184 | 0.1992 |
| theguardian.com | 0.0324 | 0.0324 | 0.0324 | 0.0143 | 0.0267 | 0.0277 | 0.0134 | 0.0143 | 0.1937 |
| who.int | 0.0170 | 0.0341 | 0.0455 | 0.0199 | 0.0085 | 0.0341 | 0.0227 | 0.0114 | 0.1932 |
| espn.com | 0.0166 | 0.0314 | 0.0300 | 0.0166 | 0.0295 | 0.0281 | 0.0244 | 0.0157 | 0.1923 |
| sentry.io | 0.0304 | 0.0292 | 0.0345 | 0.0083 | 0.0220 | 0.0339 | 0.0208 | 0.0125 | 0.1917 |
| casalemedia.com | 0.0245 | 0.0353 | 0.0299 | 0.0027 | 0.0217 | 0.0326 | 0.0217 | 0.0217 | 0.1902 |
| rubiconproject.com | 0.0150 | 0.0350 | 0.0300 | 0.0100 | 0.0200 | 0.0450 | 0.0150 | 0.0200 | 0.1900 |
| sharethrough.com | 0.0260 | 0.0355 | 0.0311 | 0.0090 | 0.0276 | 0.0289 | 0.0189 | 0.0124 | 0.1895 |
| shipt.com | 0.0204 | 0.0333 | 0.0333 | 0.0102 | 0.0213 | 0.0306 | 0.0259 | 0.0139 | 0.1889 |
| mit.edu | 0.0306 | 0.0278 | 0.0472 | 0.0111 | 0.0194 | 0.0278 | 0.0111 | 0.0139 | 0.1889 |
| doi.org | 0.0446 | 0.0446 | 0.0268 | 0.0089 | 0.0179 | 0.0446 | 0.0000 | 0.0000 | 0.1875 |
| newrelic.com | 0.0227 | 0.0277 | 0.0336 | 0.0126 | 0.0218 | 0.0235 | 0.0210 | 0.0235 | 0.1862 |
| appsflyer.com | 0.0301 | 0.0246 | 0.0346 | 0.0112 | 0.0190 | 0.0290 | 0.0234 | 0.0134 | 0.1853 |
| hp.com | 0.0208 | 0.0290 | 0.0268 | 0.0089 | 0.0283 | 0.0312 | 0.0231 | 0.0141 | 0.1823 |
| nytimes.com | 0.0245 | 0.0304 | 0.0315 | 0.0120 | 0.0141 | 0.0283 | 0.0217 | 0.0196 | 0.1821 |
| mcafee.com | 0.0230 | 0.0340 | 0.0241 | 0.0055 | 0.0230 | 0.0318 | 0.0230 | 0.0175 | 0.1820 |
| bidswitch.net | 0.0191 | 0.0226 | 0.0347 | 0.0069 | 0.0312 | 0.0243 | 0.0208 | 0.0208 | 0.1806 |
| reuters.com | 0.0282 | 0.0328 | 0.0290 | 0.0102 | 0.0226 | 0.0260 | 0.0162 | 0.0132 | 0.1783 |
| grammarly.com | 0.0298 | 0.0315 | 0.0323 | 0.0073 | 0.0218 | 0.0258 | 0.0169 | 0.0121 | 0.1774 |
| sourceforge.net | 0.0286 | 0.0254 | 0.0351 | 0.0101 | 0.0254 | 0.0217 | 0.0180 | 0.0115 | 0.1757 |
| frontapp.com | 0.0286 | 0.0248 | 0.0315 | 0.0143 | 0.0181 | 0.0296 | 0.0134 | 0.0153 | 0.1756 |
| zemanta.com | 0.0330 | 0.0312 | 0.0260 | 0.0139 | 0.0174 | 0.0260 | 0.0174 | 0.0087 | 0.1736 |
| criteo.com | 0.0292 | 0.0354 | 0.0271 | 0.0083 | 0.0187 | 0.0271 | 0.0146 | 0.0104 | 0.1708 |
| amazon.co.uk | 0.0259 | 0.0259 | 0.0291 | 0.0110 | 0.0165 | 0.0275 | 0.0126 | 0.0204 | 0.1690 |
| bbc.com | 0.0163 | 0.0272 | 0.0380 | 0.0109 | 0.0109 | 0.0326 | 0.0163 | 0.0163 | 0.1685 |
| force.com | 0.0200 | 0.0300 | 0.0350 | 0.0050 | 0.0250 | 0.0150 | 0.0100 | 0.0250 | 0.1650 |
| salesforce.com | 0.0200 | 0.0200 | 0.0350 | 0.0100 | 0.0300 | 0.0150 | 0.0100 | 0.0250 | 0.1650 |
| nginx.com | 0.0382 | 0.0215 | 0.0333 | 0.0104 | 0.0194 | 0.0181 | 0.0118 | 0.0104 | 0.1632 |
| bbc.co.uk | 0.0163 | 0.0272 | 0.0380 | 0.0054 | 0.0109 | 0.0326 | 0.0163 | 0.0163 | 0.1630 |
| cisco.com | 0.0281 | 0.0273 | 0.0315 | 0.0083 | 0.0174 | 0.0224 | 0.0132 | 0.0108 | 0.1589 |
| webex.com | 0.0273 | 0.0240 | 0.0323 | 0.0066 | 0.0182 | 0.0215 | 0.0124 | 0.0124 | 0.1548 |
| mozilla.org | 0.0142 | 0.0057 | 0.0540 | 0.0142 | 0.0142 | 0.0085 | 0.0227 | 0.0085 | 0.1420 |
| att.net | 0.0234 | 0.0234 | 0.0208 | 0.0026 | 0.0130 | 0.0312 | 0.0130 | 0.0130 | 0.1406 |
| hipages.com.au | 0.0172 | 0.0172 | 0.0172 | 0.0098 | 0.0196 | 0.0245 | 0.0172 | 0.0172 | 0.1397 |
| mozilla.com | 0.0142 | 0.0085 | 0.0483 | 0.0114 | 0.0142 | 0.0085 | 0.0227 | 0.0085 | 0.1364 |
| launchdarkly.com | 0.0179 | 0.0255 | 0.0242 | 0.0102 | 0.0077 | 0.0230 | 0.0140 | 0.0064 | 0.1288 |
| europa.eu | 0.0200 | 0.0100 | 0.0250 | 0.0200 | 0.0100 | 0.0100 | 0.0200 | 0.0050 | 0.1200 |
| aol.com | 0.0089 | 0.0179 | 0.0179 | 0.0134 | 0.0134 | 0.0223 | 0.0134 | 0.0089 | 0.1161 |
| apache.org | 0.0066 | 0.0263 | 0.0329 | 0.0066 | 0.0066 | 0.0132 | 0.0132 | 0.0066 | 0.1118 |
| adsrvr.org | 0.0112 | 0.0149 | 0.0205 | 0.0037 | 0.0131 | 0.0187 | 0.0075 | 0.0093 | 0.0989 |
| rlcdn.com | 0.0174 | 0.0145 | 0.0174 | 0.0000 | 0.0029 | 0.0145 | 0.0116 | 0.0029 | 0.0814 |
| atlassian.net | 0.0030 | 0.0089 | 0.0179 | 0.0238 | 0.0060 | 0.0060 | 0.0030 | 0.0030 | 0.0714 |

**Table 11: Website privacy policies ranked by the total ratio of the number of annotated parameters to the number of sentences in the policy *(continued from previous page).***

## L Visualizer Demonstration

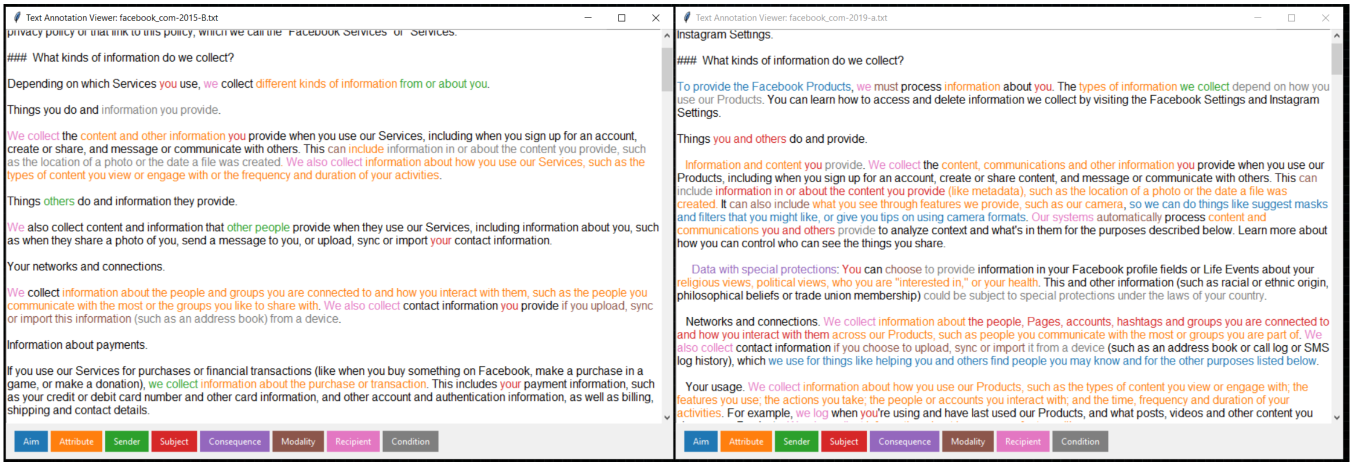


**Figure 18: Side-by-side comparison of Facebook's 2015 and 2019 policies displayed using the visualizer, showing the GKC-CI annotations in the text. The legend for the color meanings is at the bottom of each GUI window. The comparison highlights information flow changes between the policies, with parameter annotations—indicated by text color—emphasizing an increase in the specificity of information flows in the 2019 policy.**

## M Creation of Paragraph-Level Annotations

To create the paragraph-level annotations, we utilize a segmentation approach tailored to capture the structural and contextual integrity of each policy text. Instead of relying on sentence-ending punctuation to define annotation boundaries, we segmented each policy into paragraphs by identifying double newline characters. This method was chosen to reflect natural paragraph divisions in the original documents, preserving the thematic and logical flow intended by the policy authors. Following this segmentation, we proceeded with the annotation creation and benchmarking process using the same methodology used for sentence-level annotations.