common sense®

2024

Human-AI Productivity:

Evaluating Privacy Policies using Artificial Intelligence

Credits

Authors: Girard Kelly, Common Sense Media
Jeff Graham, Common Sense Media
Steve Garton, Common Sense Media

Suggested citation: Kelly, G., Graham, J., & Garton, S. (2024). *Human-AI Productivity: Evaluating Privacy Policies using Artificial Intelligence.* San Francisco, CA: Common Sense Media

Table of Contents

Abstract

The Common Sense Privacy Program was able to increase the number of published evaluations and ratings of privacy policies by developing and deploying machine learning models using natural language processing to augment and support our expert human privacy-policy reviewers. The Privacy Program's hybrid Human and Artificial Intelligence ("Human-AI") approach is able to capture high-quality annotated privacy policies by privacy reviewers. The integration of AI into our evaluation process has improved the productivity of our privacy reviewers, allowing them to complete privacy evaluations more quickly than without AI while maintaining the same or higher level of accuracy. In this paper, we examine the Privacy Program's deployment and integration of AI and attempt to quantify the cost savings and benefits to our organization and the privacy evaluation process.

Introduction

The Common Sense Privacy Program ("Program") has been addressing the privacy needs of kids, students, and families since 2016 to help them make better informed decisions based on a product's privacy practices.[1] To provide more useful information about a product's privacy practices, the Privacy Program evaluates the privacy policies of popular applications and services used by kids and families. Our privacy evaluations and ratings assist in protecting child and student data privacy, and support a more private, secure, and safer digital future for kids and families everywhere.

These evaluations of a product's privacy policy produce easy-to-use privacy ratings that help parents and educators make sense of complex practices related to popular applications and services used in the home and in classrooms across the country.[2] As of 2024, the Program has evaluated the privacy policies of over 5,000 products with expert human privacy reviewers. The Program also continually updates these published evaluations, because approximately 50% to 75% of products update their policies each year. Additionally, hundreds of companies that Common Sense has rated have worked directly with the Program's team of privacy experts to make improvements to their privacy practices, which can positively impact tens of millions of children and students across the nation. The Program also publishes privacy evaluation reports intended for policymakers and regulators, which include data and analysis of industry privacy trends over time.[3]

In this paper we examine the Common Sense Privacy Program's design, development, and deployment of Artificial Intelligence ("AI") in our privacy evaluation processes to improve scale and increase our product coverage, allowing us to communicate privacy more easily to parents, educators, and consumers without sacrificing the quality of our evaluations. The Program's work in evaluating privacy policies and providing easy-to-use privacy ratings of apps, platforms, and services is intended to empower users to make better informed decisions about privacy for themselves and with their kids and students.

Since the program's inception, increased automation, regardless of methodology, was identified as a critical path to sustainability and scale. Creating a uniform and highly structured privacy rubric enabled an evaluation process that resulted in a consistent data output that could be used to enable downstream value. Among initial automation aspects, the largest amount of time reviewers spent was on reading and identifying the relevant portions of privacy policies. In 2019, the Program began research utilizing natural language processing ("NLP") and other AI techniques as potential solutions to increase the productivity and accuracy of our evaluations of privacy policies at scale. However, unlike other approaches to privacy policy evaluation that attempt to automate and scale evaluation of hundreds of thousands of privacy policies without any human involvement or expertise, our approach was intentionally designed to combine the benefits of both artificial intelligence and expert human privacy reviewers—ensuring that there is always a human in the decision-making process.[4]

The Program has since developed, tested, refined, and deployed multiple fine-tuned AI models using the latest advancements in AI, including large language models (LLMs) and transformers. Given the timing of the development and deployment of AI, the team also created appropriate infrastructure that has crawled and classified over 300,000 privacy policies, which may include historical versions of the same policy. Our implementation is not based on recent generative AI "solutions" such as ChatGPT,[5] but is instead informed and refined using our own proprietary privacy training data produced internally by expert human privacy reviewers. Our high-quality training data is mapped to our Privacy Evaluation Framework[6] questions.

[1] Common Sense Media, Privacy Program, https://privacy.commonsense.org.

[2] Common Sense Media, Privacy Program, Privacy Ratings, https://privacy.commonsense.org/resource/privacy-ratings.

[3] Kelly, G., Graham, J., & Garton, S. (2023). 2023 state of kids' privacy: Who is monetizing our data? A general lack of transparency leads to a confusing landscape. Common Sense Media, https://www.commonsensemedia.org/sites/default/files/research/report/common-sense-media-2023-state-of-kids-privacy_0.pdf.

[4] Felin, T., & Holweg, M. (April 2024.) Theory is all you need: AI, human cognition, and decision making, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4737265.

[5] OpenAI, Introducing ChatGPT, https://openai.com/index/chatgpt.

[6] Kelly, G., Graham, J., & Garton, S. (2023). Privacy program evaluation framework. Common Sense Media, https://privacy.commonsense.org/content/resource/publications/2023-privacy-program-evaluation-framework.pdf.

Methodology

In response to overwhelming demand from parents, educators, and consumers to help them make informed decisions about privacy, the Program created a comprehensive evaluation process for applications, platforms, and online services that attempts to address some of the common barriers to understanding a product's privacy practices. The privacy evaluation process includes questions organized into categories and sections derived from the Fair Information Practice Principles (FIPPs) that underlie international privacy laws and regulations.[7] In addition, the evaluation questions and the categories that organize them are all mapped to a range of statutory, regulatory, and technical standard resources that provide background information on why each question is relevant to the privacy evaluation process.[8]

Our privacy evaluation process for an application or service is unique because it produces a score based on both transparency and qualitative details, which are combined into an overall score. These two metrics allow for an objective comparison between applications and services based on how transparent their policies are in explaining their practices, and on the qualitative nature of those practices. The privacy evaluation process was designed to summarize numerous complex privacy practices disclosed in a product's privacy policies to various audiences that include parents, educators, and consumers to support them making an informed decision using a familiar comparative format such as a product's nutrition label. In addition, our privacy reviewers are required to consistently evaluate privacy policies as accurately as possible. For example, the following evaluation question from our privacy evaluation framework requires a reviewer to read the policies of the application or service and determine whether the policies disclose the issue raised in the question by providing a yes or no response:

Question: Do the policies clearly indicate whether or not the company collects personally identifiable information (PII)?

If the reviewer responds "yes" to this question in our policy annotator software, that means the application or service discloses whether it collects personally identifiable information.[9] If there is a "yes" response to this transparency question, the reviewer is then asked a follow-up question—a slightly adjusted version of the original attempting to capture if the company or product engages in a particular practice. In this case:

Do the policies indicate the company collects personally identifiable information (PII)?

A "yes" or "no" response indicates whether personally identifiable information is, or is not, collected and will determine the final question points, based on whether the practices described are considered qualitatively "better" or "worse" for the purposes of our evaluation process. Note that some questions do not have a qualitative component. This includes questions where there is truly no qualitative value to a response, as well as questions where determining if a given response is qualitatively "better" or "worse" may require additional context outside the scope of the evaluation process. This question process may seem slightly redundant, having both a transparency and a qualitative component, but separating the disclosure from the actual practice provides an important metric for which portions of a policy may discuss a given practice. This distinction has proven valuable especially for products with overly complex or contradictory terms where identifying the actual practice presents a separate challenge. All too frequently, confusing or contradictory explanations occur in privacy policies as companies attempt to balance transparency against the urge to limit compliance liability. This tension often results in vague or contradictory privacy policy disclosures for which even trained human privacy experts and attorneys cannot agree on a shared meaning.[10]

Related Work

Other privacy-policy assessment tools and academic research into the evaluation of privacy policies at scale have explored the use of keyword-based contextual methods that attempt to summarize a policy's main issues based only on transparency. These keyword- or pattern-based methods claim to have evaluated tens of thousands or millions of products' privacy policies without the involvement of expert human reviewers. Research into privacy-policy analysis techniques indicates a wide range of advantages and disadvantages in terms of cost, efficiency, accuracy, deployment, and training. Keyword-based or analytic tools for privacy policies such as Privee,[11] PrivacyCheck,[12] Terms of Service;

[7] Federal Trade Commission (FTC). (2000). Privacy online: Fair information practices in the electronic marketplace, https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000.pdf.

[8] Kelly, G., Graham, J., & Garton, S. (2023). Privacy program evaluation framework. Common Sense Media.

[9] Common Sense Media, Privacy Program, Policy Annotator, https://policy-annotator.commonsense.org.

[10] Reidenberg, J., et al. (2014). Disagreeable privacy policies: Mismatches between meaning and users' understanding, https://lawcat.berkeley.edu/record/1126239?v=pdf.

[11] Zimmeck, S., & Bellovin, S. (2014). Privee: An architecture for automatically analyzing web privacy policies. 23rd {USENIX} Security Symposium ({USENIX} Security 14) pages 1–16, https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-zimmeck.pdf.

[12] Zaeem, R., & Barber, K. (2021). A large publicly available corpus of website privacy policies based on DMOZ. Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, https://doi.org/10.1145/3422337.3447827.

Didn't Read, [13] the Usable Privacy Policy Project, [14] PolicyLint, [15] and Polisis [16] have all been found to produce reliable measures of transparency information about key practices disclosed in an application's or service's policies. For example, previous research into automated attempts at evaluating privacy policies at scale using keyword- or artificial intelligence-based privacy policy analyses may indicate the disclosure of practices related to the terms "targeted" or "personalized" advertising, but are unable to qualitatively differentiate whether the product does, or does not, display targeted ads, or to which users, or in which scenarios. Our evaluation process was informed by this prior work through our focus on transparency, but also in identifying the absence of any clear or substantive details about specific practices.

Additional research into answering questions about privacy policies also includes data set corpora for training and includes the PrivacyQA Project, which developed a corpus consisting of 1,750 questions about the contents of privacy policies, paired with over 3,500 annotations. [17] Other privacy-policy data set corpus include MAPP, [18] which is a data set that contains 64 Google Play Store app privacy policies, and OPP-115, [19] which contains the same data types as MAPP but with 115 annotated policies. APP-350 [20] is another large data set with over 350 privacy policies with annotations on data collection and sharing. In addition, our team explored research and benchmarks of generative AI-based privacy assistant chatbots such as BingAI, Bard, and ChatGPT-4 to answer general questions about a product's privacy policies. [21]

Researchers claim these chatbot assistants, which use large language models (LLMs), provide accurate summarization of data type collection and sharing practices in a privacy policy through prompt engineering, but they suffer from the same problem of "hallucination" or confabulation that appears in other LLM models, in which these chatbots confidently provide non-substantive or incorrect answers to questions, especially if the user's prompt is not sufficiently formatted. [22] LLM-based chatbot assistants have been used to explore the analysis of legal documents and identification of relevant clauses in various types of contracts based on the user's specific request. [23] Researchers claim that techniques such as retrieval-augmented generation (RAG) produce useful case law summarizations and generalizations of legal documents that are trained on domain-specific knowledge bases, but they suffer from the same problems of LLMs in that they regularly provide false information or hallucinations. [24] Generative AI services such as Robin.Ai [25], which leverages Anthropic's [26] Claude foundation model, and Harvey.AI [27], which leverages OpenAI's GPT-4 [28] foundation model, are used in the legal professional services industry for contract summarization and review. Similarly to previous keyword- or pattern-matching AI techniques, these services do not qualitatively evaluate the meaning or practices associated with the clauses they identify for human review. These approaches focus on improving productivity by minimizing the time required for an employee to read an entire legal document, and they claim to save time by labeling the most likely relevant sections of the document for the user to make their own substantive changes based on their own specific needs and context.

Privacy Evaluation Process

Our privacy evaluation process is similar to related work in legal AI services, given that our policy annotator software similarly identifies relevant policy text from a product's policies. However, unlike recent LLM generative chatbot approaches like ChatGPT that require the user to ask a question in hopes of locating relevant sections within the contract based on the context of the question, our AI approach does not prompt the user to ask a question about a product's privacy policy, because we already know what specific questions we want to ask. Consequently, in our AI approach the question context and

[13] Terms of Service; Didn't Read, https://tosdr.org.

[14] Sadeh, N, Acquisti, A., Breaux, T., Cranor, L, McDonald, A., Reidenberg, J., Smith, N., Liu, F., Russell, N., Schaub, F., et al. (2013). The Usable Privacy Policy Project. Technical Report CMU-ISR-13-119. Carnegie Mellon University. https://usableprivacy.org.

[15] Andow, B., Mahmud, S., Wang, W., Whitaker, J., Enck, W., Reaves, B., Singh, K., & Xie, T. PolicyLint: Investigating internal privacy policy contradictions on Google Play. In 28th USENIX Security Symposium (USENIX Security 19), pages 585–602. USENIX Association, (2019). https://taoxie.cs.illinois.edu/publications/usenixsec19-policylint.pdf.

[16] Harkous, H, et al. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. https://arxiv.org/pdf/1802.02561v1.

[17] Ravichander, A., Black, A.W., S. Wilson, Norton, T.B., & Sadeh, N.M. (2019). Question answering for privacy policies: Combining computational and legal perspectives. https://arxiv.org/abs/1911.00841.

[18] Arora, S., et al. A tale of two regulatory regimes: Creation and analysis of a bilingual privacy policy corpus. (2022). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5460–5472, https://aclanthology.org/2022.lrec-1.585.pdf.

[19] Wilson, S., Schaub, F., Dara, A., et al. The creation and analysis of a website privacy policy corpus. (2016) In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 1330–1340, *Association for Computational Linguistics.* https://www.usableprivacy.org/static/files/swilson_acl_2016.pdf.

[20] Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J., Russell, N., & Sadeh, N. MAPS: Scaling privacy compliance analysis to a million apps. (2019). *Proc. Priv. Enhancing Tech*., 2019:66. https://petsymposium.org/popets/2019/popets-2019-0037.pdf.

[21] Hamid, A., Samidi, H., Finin, T., Pappachan, P., & Yus, R. GenAIPABench: A benchmark for generative AI-based privacy assistants. (2023). https://arxiv.org/abs/2309.05138.

[22] Rodriguez, D., Yang, I., Del Alamo, J., & Sadeh, N., Large language models: A new approach for privacy policy analysis at scale. (May 2024). https://arxiv.org/abs/2405.20900.

[23] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. (2024). Large legal fictions: Profiling legal hallucinations in large language models. https://arxiv.org/abs/2401.01301.

[24] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C., & Ho, D. (2024), Hallucination-free? Assessing the reliability of leading ai legal research tools. (2024). https://arxiv.org/abs/2405.20362.

[25] Robin.AI. https://www.robinai.com.

[26] Anthropic, Claude. https://www.anthropic.com.

[27] Harvey.AI. https://www.harvey.ai.

[28] OpenAI, GPT-4. https://openai.com.

scope is assumed, and we map the supporting annotated evidence in a policy to the respective question from our Privacy Evaluation Framework.[29] With this human-AI assisted approach, after a privacy reviewer selects an evaluation question, relevant AI policy annotations are automatically presented to the reviewer based on the scope of each specific evaluation question, in order to minimize the amount of policy text that a reviewer needs to read and comprehend. This approach also removes many safety concerns present in other generative AI systems that can confidently provide incorrect answers, or misleading or incomplete summarization of questions.[30] LLMs and other generative AI tools such as ChatGPT are designed not to produce truth or falsehood, because their inaccuracy is not due to misperception or hallucination. Rather, these tools lack concern with the truth and have an indifference to verifiability and empiricism, and therefore are not trying to convey information at all.[31]

With this understanding of the limitation of LLMs and generative AI, our approach focuses on utilizing LLMs to gain statistical insight into privacy policies, and it limits AI to identifying likely relevant portions of policy text and providing recommendations to the privacy reviewer, who must confirm that the AI-suggested annotation(s) are relevant to answer the question. The reviewer must then answer the corresponding qualitative component if sufficient detail is provided about a practice. The continued expert operation by trained privacy reviewers provides a continual pipeline of high-quality machine-generated and human-augmented training data. If our AI automation insufficiently identifies relevant portions of the policy, then the privacy reviewer may also manually add annotations to the process, just like they did prior to any AI integration. With this approach, the answers to each evaluation question indicate the qualitative meaning of the annotated policy text and capture additional policy text that may be used to further train or refine our AI models via supervised fine-tuning.

Privacy policies are primarily written for privacy experts and lawyers for compliance purposes, and often use complex, contradictory, or confusing language to obfuscate particular uses of personal information. Our research has found this is likely because companies simply want to minimize their legal and compliance risk, and do not want to lose customers who could learn that use of the product means their data will be monetized for the benefit of the company.[32] Understanding and coming to consensus concerning the language in privacy policies, which may be interpreted differently based on the context in which it is used, is a difficult task even for expert human reviewers. Subsequently it is not reasonable to expect AI with natural language processing to outperform expert human reviewers in this problem space without a creative hybrid human-AI approach that supports the experience and intelligence of subject matter expertise of human privacy reviewers. With this limitation in mind, our program's privacy evaluation question-and-answer process is analogous to a student taking an open book exam, where the student is evaluated on their overall reading comprehension of the content in the book through the use of several fill-in-the-blank questions.

As with an open book exam, our privacy reviewers may refer back to the relevant passages in the product's privacy policy to help them answer each fill-in-the-blank question, but they are expected to have first read the entire policy. Our human privacy reviewers are considered "privacy experts" compared to the average general population, and are required to complete our privacy training program course, which trains them on how to read and evaluate policies across multiple products at a high level of consistency and quality, while navigating the complex intersection of privacy, security, safety, and compliance issues.[33] Similar to any student taking a standardized exam, privacy reviewers have a wide range of experiences and educational backgrounds, as well as corresponding differences in ability to quickly navigate privacy policies across multiple concepts. As a result, there is some expected variation in the amount of time each reviewer spends on a particular question. The two primary variables in productivity with our human privacy evaluation process are:

- the time required to read and comprehend a product's privacy policy; and
- the time required to answer each evaluation question.

The longer a privacy policy is, the greater the overall amount of time required for a human privacy reviewer to complete a privacy evaluation. Time spent is influenced by the estimated number of pages, the approximate reading time, the reading grade level, and the number of evaluation questions to be answered. Before a product's privacy policy can be evaluated, the text of the policy URL needs to be crawled by our automated web crawling software tools,[34] then the HTML is converted to a markdown format,[35] which is both machine- and human-readable and allows highly structured annotation output using text-based offsets so we can find the source information in context if necessary. This markdown format can then be integrated into our policy annotator software to allow for human privacy reviewers

[29] Kelly, G., Graham, J., & Garton, S. (2023). Privacy Program Evaluation Framework. Common Sense Media.

[30] Rodriguez, D., Yang, I., Del Alamo, J., & Sadeh, N., Large language models: A new approach for privacy policy analysis at scale. (May 2024). https://arxiv.org/abs/2405.20900.

[31] Hicks, M.T., Humphries, J., & Slater, J. (June 2024.) ChatGPT is bullshit. *Ethics Inf. Technol.* 26, 38 (2024). https://doi.org/10.1007/s10676-024-09775-5.

[32] Kelly, G., Graham, J., & Garton, S. (2023). 2023 state of kids' privacy: Who is monetizing our data? A general lack of transparency leads to a confusing landscape. Common Sense Media.

[33] Common Sense Media, Privacy Program, Privacy Evaluation Framework.

[34] *See* Puppeteer, https://pptr.dev, which is the foundation for our headless Chrome browser policy crawler. We also use ghoststcript. *See also*, https://www.ghostscript.com for PDF policies.

[35] A custom semantically informed parser based on parse5, https://parse5.js.org, which converts HTML markup into markdown.

to use annotation tools to highlight and associate policy text for specific evaluation questions.

Crawling publicly available policies on webpages at scale and converting them into useful plain text for human privacy reviewers is not a trivial process, and has been previously attempted with limited successes by other research projects.[36] However, unlike a traditional movie or book review in which evaluation and rating occur only once, a product's privacy policies often change, sometimes more than once a year, which requires our automated crawling tools to recrawl policies periodically to ensure that each product's policies are updated to reflect any changes. To date, the Privacy Program has crawled and created AI-annotated suggestions for eight models over 300,000 policy instances, which includes historical versions of the same policy.

As policies change, we need to have insight into whether any changes in policy text are substantive for our evaluation process. We have instituted several automated procedures that help us make that determination. First, as we crawl new policies, we also periodically check our published evaluations against any new policies that have occurred to audit our question annotations and ensure that we can find the same supporting annotation. If an existing annotation for a question cannot be found in the new policy, we mark specific high-profile questions, such as the "Effective Date," to provide additional information that indicates that changes in a policy may be substantive and have impacted the annotations for that question and evaluation. Additionally, we provide a comparison of the previously evaluated policy next to the updated policy via an open-source program called wdiff,[37] which helps indicate word-by-word comparison of changes to help identify if any evaluation question responses may need to be updated. These signals help us to determine if changes to a product's policies are likely to be substantive, meaning that they may impact the evaluation's previous privacy rating, or that the policy changes directly impact our existing annotated policy text associated with one or more evaluation questions. Using these signals, our team prioritizes updated product policies for reevaluation.

Privacy Policy Readability

Readability is a reader's ability to comprehend the language used in a document, such as a privacy policy or terms of use. This is directly applicable to the ability of an individual to read and comprehend the privacy practices of a product's policies in order to make an informed decision to use the product themselves, or with their children or students. In calculating the readability of privacy policies, we use a combination of factors. For reading time, we use a custom algorithm, informed by prior readability research, that is based on a policy's text length, with an average human reading speed of 1,000 characters per minute (cpm),[38] reduced by 10% for reading on a computer screen, and an additional 10% adjusted for reading technical legal language, arriving at a rough estimate of 800 cpm. To obtain the estimated minutes required to read a policy, we divide the text length in characters by 800 cpm. We also calculate reading level by using the Flesch–Kincaid grade level algorithm.[39] Previous research into reading privacy policies has estimated the amount of time it takes to read privacy policies to only 10–20 minutes (approximately one minute per page).[40]

However, we believe this is a significant *underestimate* of the time needed to comprehend and accurately understand the various topics and practices covered in a privacy policy.[41] Similar to skimming or reading a book chapter too quickly, readers will likely only be able to answer simple or generalized questions about the book chapter's main plot and characters. Without taking notes when reading, or having specific learning goals, readers using a computer screen can often disassociate and not recall any of the smaller or more nuanced details they considered trivial when reading, such as the different characters' names or outfits, their relationships, or the names of different towns or locations that would be required for exam-level comprehension.[42] Similarly, our privacy evaluation process requires human reviewers to read a product's privacy policy with exam-level comprehension, which often requires reviewers to take notes of all the complex issues presented in a privacy policy. To complete our privacy training, and after reading a product's privacy policy, human reviewers are expected to consistently answer at least two dozen evaluation questions with a high level of accuracy. Our privacy evaluation process requires human reviewers to read policy text more slowly, intentionally, and with the precision and accuracy necessary to answer the respective number of questions depending on the type of evaluation. Our privacy research has found that the majority of privacy policies require human privacy reviewers to spend approximately 30 minutes to read a product's

[36] The Terms of Service Tracker. https://tosback.org.

[37] GNU Wdiff. https://www.gnu.org/software/wdiff.

[38] Trauzettel-Klosinski, S., & Dietz, K. (2012). Standardized assessment of reading performance: The new International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science*. 53 (9): 5452–6, https://iovs.arvojournals.org/article.aspx?articleid=2166061.

[39] Kincaid J.P., Fishburne R.P. Jr., Rogers R.L., & Chissom B.S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *Research Branch Report* 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

[40] McDonald, A., & Cranor, L. (2009). The cost of reading privacy policies. https://www.technologylawdispatch.com/wp-content/uploads/sites/26/2013/02/Cranor_Formatted_Final1.pdf.

[41] Kelly, G., Graham, J., Bronfman, J., & Garton, S. (2021). 2021 state of kids' privacy. Common Sense Media, https://www.commonsensemedia.org/sites/default/files/research/report/common-sense-2021-state-of-kids-privacy_0.pdf, pp. 41, 242 (indicates a clear majority of products require over 50 minutes of reading time).

[42] Baughan, A., Zhang, M., Rao, R., Lukof, K., Schaadhardt, A., Butler, L., & Hiniker, A. (2022). "I don't even remember what I read": How design influences dissociation on social media. https://dl.acm.org/doi/pdf/10.1145/3491102.3501899.

privacy policy for exam-level comprehension (approximately three minutes per page).

Taking into account the difficulty and complexity of reading privacy policies, our privacy training process involves systematic and rigorous coursework designed to teach non-experts how to read and comprehend privacy policies through multiple rounds of grading and feedback with an expert instructor. After the initial introduction to the evaluations and logistics of account creation and logging into the policy annotator tool, a prospective reviewer is presented with the "Basic" training question set and must complete two calibration evaluations of example products, with a focus on clarifying and standardizing the meaning and intent of the questions, as well as developing an understanding of the legal statutes and best practices behind each question. Special consideration is given to contradictory or unclear language, because the reviewer sometimes needs to "wrestle" with the transparency component (i.e., *can this question be answered*) and the qualitative component (i.e., *whether or not the practice happens*) for each evaluation question. Reviewers also need to navigate non-standard and contradictory language in policies, which can make the process difficult to calibrate across reviewers. After the calibration evaluations have been completed to a high level of accuracy, reviewers have completed our privacy training program and are able to complete evaluations to be published by a second quality assurance (QA) human reviewer to ensure standardization and accuracy. This means that every published evaluation will have been seen by at least two experts who look at the answers to each question and the supporting annotated evidence to ensure we maintain and increase the overall quality and accuracy in our published privacy ratings.

Privacy Evaluation Types

The Privacy Program completes three different types of privacy evaluations using human privacy reviewers to meet the unique requirements and needs of different audiences that include parents, educators, consumers, policymakers, and experts. In order to help these diverse audiences make smart choices about privacy for themselves and their children or students based on their specific needs and concerns, we created three types of privacy evaluations. These types of evaluations are differentiated by the number of privacy evaluation questions required to be answered, which are segmented by their intended audience, and are labeled "Quick," "Basic," and "Full" privacy evaluations:

- **Quick Evaluations**. The majority of products intended for kids and families receive a Quick evaluation, which is meant to help parents, educators, and consumers *make quick decisions about privacy*. Quick evaluations produce a privacy rating that is meant to help answer one of the most common privacy questions that parents and educators ask about a product, which is: "Is the company making money from my data?" This type of rating is based on only seven unique evaluation questions, and is sufficient to determine our "Pass," "Warning," or "Fail" rating based on answers to the rating questions. The seven Quick question topics cover a policy's minimum privacy compliance obligations and use of personal information for commercial purposes, which include: 1) the *Effective Date* of the policy; 2) *Third-Party Marketing* communications; 3) *Selling Data* to third parties; 4) displaying *Personalized Advertising*; 5) using *Third-Party Tracking* technologies; 6) *Tracking Users* across sites and services; and 7) creating *Advertising Profiles*.[43]
- **Basic Evaluations.** The most popular products intended for kids and students receive a Basic evaluation, which is a more comprehensive evaluation meant to help parents, educators, and consumers make basic informed decisions—beyond data monetization—about the different privacy issues that matter most to them. Basic evaluations are more comprehensive than Quick evaluations because they help answer the most important privacy questions about a product across 10 different *Evaluation Concern* categories. This type of evaluation is based on 28 Basic evaluation questions that include the seven Quick evaluation questions. Similar to Quick evaluations, Basic evaluations also display a rating. In addition, Basic evaluations display an *Overall Score* to more easily compare different products across multiple privacy issues for use in different contexts.
- **Full Evaluations**. The most popular Big Tech products used by millions or hundreds of millions of people receive an extensive 155-question inspection of all the possible privacy and security evaluation questions about a product, including all questions covered in our Quick and Basic evaluations. The Full evaluation is our most comprehensive evaluation that we provide and is meant to help parents, educators, consumers, experts, researchers, and policymakers make the most informed decisions possible about all the different types of privacy issues that matter most to them. Only a limited number of Full evaluations are completed by our reviewers for research purposes each year.

Building a Better Privacy Nutrition Label

Quick evaluations are the fastest for human privacy reviewers to complete, but they only produce a privacy rating (Pass, Warning, Fail) based on the most important seven evaluation questions that cover use of personal information for commercial purposes. Our Program started our AI integration with only the seven privacy rating questions in order to validate that our AI

[43] Common Sense, Privacy Program, Privacy Ratings. https://privacy.commonsense.org/resource/privacy-ratings.

approach can scale while not sacrificing the accuracy or quality of our privacy ratings. The next step and challenge will be to expand the number of AI-enabled questions from Quick-7 to Basic-28. This will support all of our audiences in making better-informed decisions about privacy thanks to the presence of a privacy rating, privacy score, and more detailed information across 28 rather than seven questions. This will also create rich, structured data that will enable the exploration of better options to communicate privacy practices in a consistent and easily understood manner. The Quick evaluation is limited to only seven questions, and therefore only provides a quick overview of a few high-concern details because the scope of questions is not broad enough to provide a deeper level of privacy analysis.

Basic evaluations with 28 questions that represent all of our evaluation categories capture a more realistic minimal evaluation that enables communication across a diverse range of privacy, security, safety, and compliance practices. The ability to reach different audiences that have different levels of knowledge when it comes to privacy requires an additional level of detail and nuance to meet their context and concerns.[44] Similar to comparing ingredients of food products, the additional data provided by Basic evaluations is important because consumers, parents, and educators are more able to compare the different privacy practices of a product to other products in order to make an informed decision based on their specific needs. Our audiences often describe our privacy ratings as a simple product assessment of "healthy" or "unhealthy," and so our Basic evaluations are used to break down the product's practices into the most important ingredients like those displayed on nutrition labels, such as calories, fat, sugar, carbs, or allergens.

Privacy policies are often long and difficult to read, but—just like nutrition labels—they are also a critical part of product transparency that informs users about the data that the product collects and the promises a company makes about how they use that data. The more transparency a company provides in its policies and labels about its product's privacy practices, the more information parents, educators, and consumers have to make better choices for themselves, their children, and their students. Clearly some nutritional details are more important to some consumers than other details, such as whether the product contains wheat, dairy, nuts, or meat, given that the shopper may have dietary restrictions or allergies. Analogously, when thinking about privacy practices, a lack of transparency for an application or service may present different challenges relative to the contexts in which it could be used, such as at home, in the classroom, at work, or in public places. In addition, products may be used by different audiences with different needs that may require different accommodations and protections, such as consumers, parents, educators, or their children or students.

[44] Common Sense, Privacy Program, Evaluation Details. https://privacy.commonsense.org/resource/evaluation-details.

Just as regulators control our food system to protect the public's health and the safety of food products in the grocery store, we also need to protect the privacy of kids and families based on the product's actual practices. A product's privacy practices, just like nutrition labels, need to be explained consistently across products using simple, clear, and easy-to-understand language. When a product's features and practices are accurately disclosed, consumers can make informed decisions about whether to purchase or use a product, which may build consumer trust and brand loyalty, resulting in increased product demand over time. Contrary to some popular belief, consumers have acquired the general skills and knowledge to navigate the current digital marketplace of goods and services, despite countless features and prices, to choose the products they need.[45] They know what factors are the most important to them in making an informed decision to purchase a product. Because of the limited time they have to make a decision, reading lengthy privacy policies does not allow consumers to quickly and confidently identify the most important privacy practices of a product necessary to compare and contrast similar products.

Historically, basic food nutrition labels have served the purpose of educating consumers about the most important nutritional details or ingredients of a product, and as a result are highly regulated to protect the public's health and safety. Nutrition labels have created consistency and trust in the competitive marketplace, where consumers and companies both have the same expectation of what a nutrition label should include, and they have settled on the language and terms that consumers need to know in order to differentiate products in the marketplace. When evaluating products with food nutrition labels, consumers often use the amount of various nutrients, in grams or as a percentage of their daily intake, to make an informed decision based on their specific needs. At a high macro level, those details include calories, carbohydrates, and fat, but for some people, especially those with life-threatening or serious allergies, allergens are a very important detail, and these are typically listed in the ingredients section. And there has been a move to explicitly list some of the top allergens (e.g., peanuts, milk, eggs, soy, wheat, shellfish).

Our Basic privacy evaluation questions are like a list of ingredients in that they cover a wide range of unique issues that make up a comprehensive landscape of all the FIPPs

[45] *See* Perrin, A. (April 2020). Half of Americans have decided not to use a product or service because of privacy concerns. Pew Research Center, https://www.pewresearch.org/short-reads/2020/04/14/half-of-americans-have-decided-not-to-use-a-product-or-service-because-of-privacy-concerns. *See* also Auxier, A., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). Americans and privacy: Concerned, confused, and feeling lack of control over their personal information. Pew Research Center, https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information.

privacy principles[46] across privacy, safety, security, and compliance-related concerns that would reasonably be expected to be disclosed in the privacy policies and labels of products intended for children, students, and families. However, privacy policies often do not disclose what data types they **do not** collect, or privacy practices they **do not** engage in, even when those practices are important factors when comparing products and helping make an informed decision. If a product's food nutrition label **does not** disclose that it contains a harmful ingredient consumers look for when making a decision on whether to purchase the product—such as wheat, dairy, or nuts—that means the product **does not** contain those potentially harmful ingredients, and therefore doesn't pose a risk for that particular consumer.

Alternatively, products are proud to indicate they contain "zero fat" or are "low calorie," clearly showing that consumers are motivated by the absence of some ingredients. Similarly, consumers want to use products that **do not** sell their data and **do not** compromise their privacy. But there is no baseline expectation across all products about what a product may or may not do with data, and therefore the absence of details in a privacy policy about a particular practice **does not** mean that the practice isn't occurring. This is confusing to consumers.

There are certainly challenges with nutrition labels, such as when companies cannot agree on a standardized definition of what a product contains (eg. "may include dairy"), or even on the ingredient's origin (eg. "may include oranges from Mexico, USA, or Brazil"). Nutrition labels may also use multisyllabic and misleading terms to describe ingredients.[47] We see similar issues in the privacy landscape with obfuscating language. In some cases, the vocabulary and societal awareness is lacking that is necessary to discuss these issues with consistency.[48] This lack of consensus results in the use of different language to explain similar privacy practices, which further confuses consumers. A product's privacy policy and privacy rating that **does not** disclose its "worse" practices means the product should not be presumed safe, because the product still reserves the right to engage in the "worse" practices without any notice, thereby putting children and students at risk for potential harm. As an example, if a product's policies don't mention any practices regarding selling data, it's not safe to assume that the product doesn't sell data.

Regardless, for food labeling, some ingredients are so harmful to some people that top allergens are often explicitly disclosed for clarity (eg. Does Not Contain: Nuts, Dairy, Wheat, Soy). Many privacy issues pose similar risks and should be explicitly disclosed regardless of whether a product engages in the practice.[49] Unlike a product's nutrition label, a company's privacy policy that is not clear or transparent regarding a privacy practice means the product may still engage in that practice without providing any notice. Therefore, privacy policies should always disclose **whether or not** they engage in the most important practices consumers, parents, and educators want to know about concerning their privacy. The Basic evaluation questions were designed with nutrition labels in mind to align as closely as possible with transparency and the expectations of consumers, parents, and educators who all have different expectations. These audiences need as much information as possible about a product's privacy practices to make an informed decision.

Privacy Evaluation Questions & Answers

Our privacy evaluation process breaks down how our reviewers evaluate privacy policies based on reading them, so that each component is optimized for productivity and consistency. The four-part, multi-step question and answer process assumes prior experience and training with our methods, our questions, and the privacy policies for the products being evaluated. The process for answering one question is shown below, with an estimated time for each step. Note that we break down the time estimates based on whether a product is transparent about a given practice. Even if a product doesn't discuss a topic, our process requires the reviewers to confirm the *absence* of substantive details about that practice. Capturing the lack of transparency reveals a subtle but valuable detail that's often missing from other privacy assessment frameworks, where often the lack of details about a topic may lead to the incorrect assumption that more privacy-protecting practices are in use than there actually are.[50]

For a Quick evaluation, when a privacy policy is transparent for a question, the process looks like the following and typically takes approximately 4–6 minutes:

- **Question (30 seconds):** The reviewer must read and comprehend each privacy evaluation question asked, just like a multiple-choice exam question.

[46] Federal Trade Commission (FTC). Privacy online: Fair information practices in the electronic marketplace. (2000). https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000text.pdf.

[47] Jacobs, A. (2021). Lawsuits over 'misleading' food labels surge as groups cite lax U.S. oversight. *New York Times*). https://www.nytimes.com/2021/09/07/science/food-labels-lawsuits.html.

[48] Kelly, G., Graham, J., & Garton, S. (2023). 2023 state of kids' privacy: Who is monetizing our data? A general lack of transparency leads to a confusing landscape. Common Sense Media. (The distinction between sell and share can also confuse consumers who do not assume "sharing" their data with third parties is the primary method that companies use to monetize their data, because the term "sharing data" predates the CCPA or CPRA's appropriation of "share.")

[49] Kelly, G., Graham, J., Bronfman, J., & Garton, S. (2019). Privacy risks and harms. Common Sense Media. https://privacy.commonsense.org/resource/privacy-risks-harms-report.

[50] For more background on why a lack of transparency in privacy policies is typically hiding invasive practices, see Kelly, G., Graham, J., Bronfman, J., & Garton, S. (2021). 2021 state of kids' privacy. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/common-sense-2021-state-of-kids-privacy_0.pdf.

- **Transparency (30 seconds):** The reviewer must use their memory or prior recollection of reading the product's privacy policy to determine whether the policy transparently disclosed information related to the issue presented in the question.
- **Qualitative (1 minute):** If the reviewer determines that the policy does adequately disclose a specified practice presented in the evaluation question, then they must also determine from memory whether or not the product engages in the specified practice.
- **Annotations (2 minutes):** Finally, if the reviewer determines that the policy transparently discloses the specified practice, they must provide supporting evidence from the policy to justify their answer by annotating a sufficient amount of sentences, clauses, or paragraphs.

When a privacy policy is non-transparent for a given question, the process looks like the following and typically takes approximately 2 minutes:

- **Question (30 seconds):** The reviewer must read each privacy evaluation question, just like a multiple-choice exam question.
- **Transparency (30 seconds):** The reviewer must use their prior recollection of reading the product's privacy policy to determine whether the policy transparently disclosed information related to the issue presented in the question. If the policy is non-transparent on the issue, then the transparency answer to the question is "No," and the reviewer moves on to the next evaluation question.
- **Verification (1 minute):** If a reviewer indicates an evaluation question is non-transparent, the reviewer is expected to spend 1–2 minutes reviewing the policy to ensure that it doesn't disclose any other information related to the question, for example by possibly using non-standard or contradictory language that was not captured by our AI model.

Table 1 indicates that this multi-step process for each question could be further optimized to reduce the time it takes to determine whether a policy is transparent about a practice, to ascertain what that practice actually is, and to locate and annotate evidence from the policy to justify the answers. The process also allows for the capturing of expert human privacy reviewer knowledge in evaluating privacy policies. We used this knowledge in our initial AI training data.

It typically takes our privacy reviewers four to six minutes to complete each evaluation question, as described above. Privacy reviewers may take less time to answer questions with clear and concise transparent disclosures, but will often take longer on more complex or difficult questions that may have confusing or contradictory language. Given the complex nature of reading, comprehending, and evaluating privacy policies, our privacy reviewers indicate most questions are considered difficult, and reviewers need to "wrestle with the policy" to understand the context of disclosures relative to our questions and correctly answer each question. In order to capture details that may fall outside of the described standardized process, we provide two tools that reviewers may use to document any abnormalities. First, reviewers can use the flag feature, which can signal something notable, non-standard, or requiring secondary review or confirmation. Second, we provide an open-ended notes text field where reviewers may leave additional details about why a question is answered in a specific way—especially useful when policies include contradictory or ambiguous language, and the answer serves as the final judgment of whether the practice takes place. These extra data fields provide valuable context and insight in the event we are ever asked to justify our privacy rating, or for a future quality assurance reviewer who may be updating a product's evaluation, or if we are asked to speak about specific answers to questions.

If a question is non-transparent, privacy reviewers take on average one minute to complete the first two sub-steps and are expected to move onto the next question, but may spend one additional minute reviewing the policy to ensure they didn't accidentally miss other disclosures that could address the issue. This is sometimes necessary when policies are confusing, contradictory, or use non-standard language to discuss a given practice. Assuming the average privacy policy length of 12 pages, it typically takes a reviewer 30 minutes to read a policy with a limited number of issues in mind if the average reading time is approximately three minutes per page.[51,52] For Quick evaluations, reviewers are expected to skip sections of the privacy policy that are likely irrelevant to answering the seven Quick evaluation questions. This results in an estimated 50% reduction in policy reading time compared to reading the entire privacy policy and additional Terms of Use. Assuming a high level of transparency across all questions, it typically takes 60 minutes for a human reviewer to manually complete a Quick evaluation, including reading the privacy policy (30 minutes) and answering the seven questions (30 minutes).

For our Basic evaluations, which include 28 questions, our experience indicates it takes reviewers longer to both read the product's policies and to answer the questions. Each sub-step takes slightly longer to complete due to the higher cognitive demand needed to process, retain, and recall an increased amount of more complex information. Our experience also indicates Basic evaluations take longer to complete per question than Quick evaluations. We assume this is due to the increased

[51] Lee, I. (2018). It's not you; privacy policies are difficult to read, Privacy Program, Common Sense Media. https://www.commonsense.org/education/articles/its-not-you-privacy-policies-are-difficult-to-read.

[52] Kelly, G., Graham, J., Bronfman, J., & Garton, S. (2021). 2021 state of kids' privacy. Common Sense, pp. 37–39,170–171. https://www.commonsensemedia.org/sites/default/files/research/report/common-sense-2021-state-of-kids-privacy_0.pdf.

Table 1: The amount of time required to complete each evaluation type. All times are approximate ranges based on our experience. Individual products and specific questions may fall well outside the listed times for various reasons such as policy length and/or contradictory or particularly confusing language.

Evaluation Type	Policy Time	Question	Transparency	Qualitative	Annotations	Time per Question	Expected Completion Time
Quick (7)	30 mins	30 secs	30 secs	60 secs	120 secs	4-6 mins	60 mins
Basic (28)	60 mins	30 secs	60 secs	90 secs	180 secs	6-8 mins	240 mins

number of questions presenting more diverse and complex issues, requiring reviewers to take more time per question to recall more information and provide more annotations. It is assumed, basic privacy reviewers take more time to complete each sub-step of an evaluation question for a total of six-to-eight minutes per question given the increased cognitive load and task switching across disparate topics.

For a Basic evaluation, when a privacy policy is transparent for a given question, the process looks like the following and typically takes approximately 6–8 minutes:

- **Question (30 seconds):** The reviewer must read and comprehend each privacy evaluation question asked, just like a multiple-choice exam question.
- **Transparency (1 minute):** The reviewer must use their memory or prior recollection of reading the product's privacy policy to determine whether the policy transparently disclosed information related to the issue presented in the question.
- **Qualitative (1.5 minutes):** If the reviewer determines that the policy adequately discloses a specified practice presented in the evaluation question, then the reviewer must also determine from memory whether or not the product engages in the specified practice.
- **Annotations (3–5 minutes):** Finally, if the reviewer determines that the policy does transparently disclose the specified practice, they must provide supporting evidence by annotating one or more sentences, clauses, or paragraphs of every occurrence in the policy that is relevant to the issue presented in the question.

Given the increased complexity, scope of issues, longer reading time, and more annotations across a product's multiple policies, it typically takes a reviewer approximately 4 hours to complete a Basic privacy evaluation. The reviewer is expected to read multiple policies that include the product's privacy policy and Terms of Use (60 mins), and answer all 28 Basic questions (3 hours).

Privacy Evaluation Cost Analysis

Table 2 indicates that each privacy reviewer earns a fixed rate of $125 (or approximately $31.25 per hour) to read, analyze, and complete a single Basic privacy evaluation of 28 questions taking approximately 4 hours. This fixed rate and hourly wage was calculated to align with the average hourly wage in 2023 for paralegals and legal assistants performing expert and complex legal analysis.[53] With a fixed cost of $31.25 per hour for each trained privacy reviewer, it costs $31.25 to publish seven questions in approximately 60 minutes. This results in an approximate cost of $4.46 per Quick question answered.

Our organizational goal is to publish Quick privacy evaluations for 10,000 products, which would mean we could provide privacy ratings for *all* products reviewed by Common Sense. At a fixed cost of $31.25 per evaluation, it would cost the Privacy Program approximately $312,000 to scale to 10,000 products and approximately 10,000 human reviewer hours—or one full-time employee working almost 5 years—to complete and publish Quick privacy evaluations for all of the most popular applications and services used by kids and families. After the 10,000 privacy evaluations are published, there will also be an ongoing evaluation update maintenance cost of approximately 60% of the cost to evaluate these 10,000 products each year. This will keep the evaluations current as products update their privacy policies with new practices.

However, when considering how our Program can scale privacy evaluations to cover 10,000 products, we should remember that it's important to our audiences that we not only provide Quick evaluations (with seven questions) that give limited information about a privacy rating, but also that we consider how we can provide more Basic evaluations (with 28 questions). This would ensure every privacy evaluation includes more nuanced details, including a privacy score, so that consumers, parents, and educators can make more informed decisions about privacy.

Table 3 indicates that with the current process, our human expert privacy reviewers could complete 10,000 Basic evaluations, including a privacy rating and privacy score, at an approximate cost of $125 per evaluation.

[53] The Bureau of Labor Statistics (BLS) reported an average annual salary of $66,460, and an average hourly wage of $31.95, for paralegals and legal assistants working in the US in 2023. Bureau of Labor Statistics. (May 2023). Occupational employment and wages, paralegals and legal assistants. https://www.bls.gov/oes/current/oes232011.htm.

Table 2: Fixed cost for each evaluation type.

Evaluation Type	Policy Time	Time per Question	Total Time	Cost per question	Cost per evaluation	Time per evaluation
Quick (7)	30 mins	4 mins	60 mins	$4.46	$31.25	1 hour
Basic (28)	60 mins	6 mins	180 mins	$4.46	$125	4 hours

This would cost the Privacy Program approximately $1,250,000 to scale, and approximately 40,000 human reviewer hours, to complete and publish Basic evaluations for all of the most popular applications and services used by kids and families. This would require more than 20 full-time human privacy reviewers to complete all the evaluations within a year and clearly would not be practical given the budget constraints of a grant-funded, small non-profit organization. As discussed, given that about 60% of products change or update their privacy policies at least once per year, there is an additional maintenance cost of approximately $187,500 annually to manually re-evaluate Quick evaluations, and approximately $750,000 annually to manually re-evaluate Basic evaluations for 10,000 products.

Privacy Trained Artificial Intelligence

Artificial intelligence, or AI, is a transformative technology capable of augmenting and supporting specific business needs and processes with powerful predictive and analytical capabilities. However, AI is not a magic solution to every problem. AI is just a set of algorithms, each with its own strengths and weaknesses that may be suitable for some types of problems, but not suitable for other problems in different contexts.

Before deciding to develop and integrate AI into the Program's privacy evaluation process, it was important to intentionally separate the hype of AI from reality, and evaluate our specific needs and requirements to determine whether use of AI would positively or negatively impact the specific problem we are trying to solve. Research has shown that approximately 80% of company projects attempting to integrate AI fail.[54] Therefore, to successfully integrate AI into our evaluation process, the Program determined exactly which business processes would benefit from the integration of AI and which processes would not. The Program also evaluated how we would collect high-quality AI training data, and how we would deploy current and future AI models into our production environment. In addition, after the strategic decision was made to develop and deploy AI in our privacy evaluation process, it was imperative that we include a feedback loop to help calibrate and appropriately integrate AI into our workflow without sacrificing quality for scale.

To determine if our AI integration was actually improving our process, we asked human experts to evaluate whether the deployment of AI was positively or negatively impacting their decision making. Additionally, we improved our Quality Assurance (QA) process to ensure that no compromises to accuracy were being made prior to publication of new evaluations. If the integration and beneficial use of AI is going to be successful, ongoing testing of AI in real-world scenarios is crucial, because the world is an incredibly dynamic system and a model that performed well yesterday may not perform well tomorrow. Intuitively this makes sense, as new privacy laws may require companies to disclose specific practices using specific language, or with a specific level of detail, and as a result the language in the policies may change even though the policies themselves may not substantively change. If AI is expected to inform critical decision making, any deployment of it should be evaluated by experts capable of assessing the performance and accuracy of the system.

The overall cost-benefit analysis of AI versus a non-AI-augmented system is certainly complex. When considering the cost of an AI system, details like the following need to be included:

- The time required to train and deploy new models
- The AI's accuracy compared against the experts it's augmenting or replacing
- The business consequences if AI makes mistakes
- Ongoing data quality
- Data infrastructure maintenance
- Expert retraining costs over time

Ultimately, we chose to integrate AI into our privacy evaluation process, and we've carefully monitored how it's used by our human privacy reviewers over the course of several years. The Program also realized the expected cost savings as well as increased productivity, all while ensuring evaluations maintained or improved the high level of accuracy previously achieved by human reviewers without the use of AI. Integrating AI into the privacy

[54] Ryseff, J., De Bruhl, B., Newberry, S. The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed. Avoiding the Anti-Patterns of AI. (2024). https://www.rand.org/pubs/research_reports/RRA2680-1.html.

Table 3: Completion costs for each evaluation type.

Evaluation Type	Number of Evaluations	Cost per Evaluation	Completion Cost	Expected Hours to Complete	Annual Update Cost
Quick (7)	10,000	$31.25	$312,500	10,000	$187,500
Basic (28)	10,000	$125	$1,250,000	40,000	$750,000

evaluation process, using natural language processing, helps support the expert reviewers by recommending relevant privacy policy text in the form of clauses, sentences, or paragraphs from the product's policies that are relevant to answering each evaluation question. This approach increases impact by automating parts of the privacy evaluation process that were tedious or that caused reviewers to spend a disproportionate amount of time on mundane tasks that did not take advantage of their rich depth of expertise.

The integration of AI into our day-to-day reviewer workflow has enabled our Program to develop efficiencies of scale, and thus to rate more products. Our AI integration is also included in our existing policy annotation software, enabling a continuous pipeline of privacy policy annotations mapped to our evaluation framework. This also provides a high-quality feedback loop that includes over 5,000 unique privacy evaluations answering Quick-7 questions, with over 1,000 of those also answering at least our Basic-28 questions. To our knowledge, this level of high-quality privacy training data is unparalleled in the industry.

The overall objective of integrating AI into our privacy evaluation process is to increase evaluation throughput and accuracy while maintaining evidence-based evaluations that provide information on more privacy-protecting product choices, as well as informing research and advocacy efforts. This is accomplished by reducing the time needed for a human reviewer to evaluate a product's privacy practices. Our use and integration of AI will always require a "human in the loop" or Quality Assurance (QA) oversight, even as the Program tries to further reduce the amount of time to evaluate policies as much as possible without a sacrifice in quality or accuracy.

In addition, the Program's nuanced privacy evaluation questions are founded on domain-specific knowledge of legal, societal, educational, and child development issues. An important differentiator of the Program's approach to using AI is to display AI-suggested annotations for each question that is selected by the human reviewer as the source of truth. Therefore, our AI models ensure that accurate, consistent, and relevant policy annotations will contribute to future model improvements as the Program scales up and rates more products. Lastly, because human reviewers include recommended relevant annotations with each evaluation question, the Program continues to use annotated evidence from the product's policy to help navigate difficult questions and disputes from companies asking about the accuracy of their privacy rating.

Initial Artificial Intelligence Research

In early 2019, the Program began exploring the capabilities of various AI approaches. The initial research did not encompass the groundbreaking work using transformers described in the paper "Attention Is All You Need,"[55] and thus we proceeded to explore various approaches of text-feature extraction[56] using scikit-learn[57] and other vectorization techniques such as GLOVE.[58] At the same time, various segmentation processes and approaches were explored using open-source projects like NLTK[59] and other segmentation approaches. We also explored rudimentary preprocessing and various approaches including CNN,[60] RNN,[61] and SVM,[62] as well as various clustering algorithms including k-means clustering. Our approaches that used these technologies did not achieve any level of performance worth considering as a potential route for saving reviewer time during the privacy evaluation process. The performance of any models we produced were simply unusable for our requirements.

Undaunted by early setbacks, we continued research and exploration until we came across transformers and started exploring early LLM models such as BERT[63] and RoBERTA[64] with the aid of early libraries like simpletransformers,[65] which leveraged Hugging Face's openly available LLMs and made the space considerably more

[55] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2023). Attention is all you need. *arXiv [Cs.CL]*. http://arxiv.org/abs/1706.03762.

[56] *See* https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction.

[57] *See* https://scikit-learn.org/stable/index.html.

[58] Pennington J., Socher R., & Manning C. (2014). GloVe: Global Vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics. https://aclanthology.org/D14-1162.pdf.

[59] *See* https://www.nltk.org.

[60] *See* https://www.ibm.com/topics/convolutional-neural-networks.

[61] *See* https://www.ibm.com/topics/recurrent-neural-networks.

[62] *See* https://www.ibm.com/topics/support-vector-machine.

[63] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv [Cs.CL]*. http://arxiv.org/abs/1810.04805.

[64] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). "RoBERTa: A robustly optimized BERT pretraining approach." *arXiv [Cs.CL]*. http://arxiv.org/abs/1907.11692.

[65] *See* https://pypi.org/project/simpletransformers.

accessible. During most of this early exploration, we spent a large amount of time exploring various hyperparameter values and their impact on the fine-tuning process to create binary and multi-modal classifiers.

As the transformer research space matured considerably in 2020, we eventually settled on using ktrain[66]—a library that uses simple abstractions and includes high-quality documentation. At this point, the early research stabilized and we were seeing promising results using LLM-backed classifiers. We decided to move forward building fine-tuned binary classifiers backed by distilbert-base-uncased[67] with one model per question. Our binary classifiers would indicate one of two possible classes, "related" or "not related," for a given policy segment and a given evaluation question. This approach was straightforward and identified relevant policy text as well as providing some, albeit limited, ability to inspect their behavior using techniques like LIME[68] to understand how they performed, especially when they did not perform well. This approach also had the benefit of easily becoming integrated into our existing evaluation process, and enabling a feedback loop for future model improvements.

When building and refining AI models, we employed F1 scores as a measurement for usefulness. It is important to note that for our intended use, accuracy would be a poor measure for the performance.[69] Typical policy text only has a tiny portion, typically less than 1%, of policy text that is relevant for a given question or concern. If we were to create a "model" that always indicated any excerpt of a policy was "not related" to a given question, our model would be over 99% accurate. The F1 score mitigates this issue by taking into account the various types of errors a classification process could make. Similarly, since our data is so skewed, we focused on the F1 scores for the "related" class only, as using the macro average F1 score would be heavily skewed by the disproportionate amount of "not related" data. In our experience, reporting the macro average F1 score would typically achieve an inflated value in the range of 10% to 40% higher than only the "related" F1 score. Therefore, our model's macro F1 scores would likely not be an accurate reflection of, or provide meaningful insight into, how the models may perform for their intended use. Please see the *Appendix* for a more detailed discussion of F1 scores.

The Road to Production-Ready

The Program created initial unrefined AI models for all Basic-28 evaluation questions to better understand what to expect and where our challenges would be in bringing these models into a production environment. This allowed us to get a baseline understanding of model performance. From our experience, the F1 score for each model was—unsurprisingly—dependent on the number of policy examples available in our training data. We also saw that for questions whose topics are backed by privacy laws that are explicit about requirements, and where vendors are transparent and use clear language, our AI models tended to perform better than they did on questions not sharing those characteristics.

Conversely, on the lower end of F1 scores, we see few policies that disclose details related to those questions, or we see privacy policies containing language that is contradictory, vague, or obfuscating. These issues also present challenges for human expert privacy reviewers to accurately classify.

We use the annotations from our expert human-evaluated privacy policies as the ground truth or "gold standard" for our supervised machine learning process. In order to refine and improve the performance of the model, we use a feedback-loop-driven iterative process in which early models are used to evaluate the training data and present examples where the model and reviewer disagreed on the classification of a given excerpt of policy text. In that case, an additional privacy expert reviews the mismatched AI suggestion and human-provided annotation. The additional domain expert is presented with the model's prediction and the reviewer's annotation, and then makes a judgment call to indicate that either the model is accurate and the annotation should be adjusted by being added to or removed, or that the reviewer is correct and no adjustment is necessary. Please see the *Appendix* section "Annotation Cleanup & Alignment" for more details on this process.

Our incoming training data is imperfect for several human-related reasons:

1. Human reviewers, while very skilled, sometimes skip providing **all** of the annotated examples for a given question. Typically humans will annotate a sufficient amount of text to accurately answer a question, but may omit redundant information or information that is related, but does not change a question response given the other existing annotations. These types of omissions save time and do not impact the response of the final question.
2. Sometimes the annotation process is imprecise and the annotation may include surrounding text that is not actually associated with a question. These types of errors are typically quite obvious to humans, but automated systems may not properly ignore the

[66] Maiya, A. S. (2004). Ktrain: A low-code library for augmented machine learning. *arXiv Preprint arXiv:2004. 10703*. http://arxiv.org/abs/2004.10703.

[67] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter." *ArXiv*, https://arxiv.org/abs/1910.01108.

[68] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. *arXiv [Cs.LG]*. http://arxiv.org/abs/1602.04938.

[69] See the Appendix for more details on F1 scores.

imprecision and include noisy information into the corpus.

Additionally, the automation details such as corpus building, data preprocessing, and other preparation processes may create additional challenges in having model predictions align with human-provided annotations:

1. There may be a mismatch between how a human performs an annotation task versus how the automated segmentation process is performed. As a result, machine-segmented policy text may not be appropriately aligned on boundaries and may result in a mis-mapping to the proper annotation, so some items in the corpus could end up mislabeled.
2. The segmentation may not properly identify appropriate boundaries to break down text, so it may include superfluous or unrelated details.
3. Our current process does not consider lead-in or surrounding context of an excerpt, so we miss context that may be available if a particular segment is analyzed in isolation. This admittedly was an impediment in early LLM models, but technological advancements have created more powerful models that have fewer limitations today. However, those larger and more robust models typically require considerably more resources to fine-tune and use. We have not yet explored whether the tradeoff of using larger and more powerful current LLMs could provide enough value relative to their increased demands on resources. Additionally, they may not be able to provide explicit source citation with a narrow enough focus to help us identify the **exact** portions of a policy that support a given response.

The process of refining and aligning our human-annotated privacy evaluations with our corpus processing and preparation makes up a considerable portion of the cleanup and alignment work we find necessary to improve model performance.

Taking the time to refine the corpus serves two purposes. First, it helps to ensure that any F1 scores we consider are more likely reflecting real-world performance, where we entirely rely on automated segmentation and other processes without human intervention. Second, it ensures we are sending more consistent information to the training or fine-tuning process. In most cases, current cleanup and refinement addresses straightforward mismatches in segment boundaries, or missed annotations by the human reviewer. As refinement continues, each generation of the model learns more patterns in the data with more nuance and in what contexts certain words like “sell” may or may not be of importance. After a few rounds, the mismatches between human-provided annotations and AI suggestions become more challenging to correct, and we need to make tougher decisions as to whether some annotations on the edge of a topic should or should not be included.

Most of our additional tooling was crafted to help align our human-provided annotations with what our automated segmentation provides. This process is imprecise and introduces some additional room for error. The annotations may have poor boundaries as created by our automatic segmentation, or the human reviewer may have been imprecise and included too a small portion of the previous or subsequent segment, which may be unrelated. To account for this error, our first pass of the mapping assumes that the human-provided annotation is intentional and that any extra or partial annotations contain important context that we should pay attention to. This means we err on the side of ensuring we include all human-provided information, even if in context it may not obviously be related to a given question. This process is entirely automated. It takes completed evaluations as input and builds an appropriate corpus ready for model training. After the initial processing to align human annotations with automated segments, our corpus builder then creates an appropriate training and testing validation set for respective questions, so that we can begin the model training process.

Ideally we at least want a model that is not *confidently incorrect*. In cases where it leans toward an incorrect classification, we would prefer ambiguity, as that provides some information about indicating that a policy excerpt may be related to a question. At that point we can let a human make the final judgment call. When models are confidently incorrect in their answers, they provide us little insight and add little value to the process. In some cases, confidently incorrect models could possibly even make things worse as they may perform incredibly well in other scenarios, leading reviewers to trust the model's answers more than they should. [70]

Independent research into label classification and use of predicted answers over time has found lowered quality in the performance and accuracy of human reviewers, because the reviewers may place unwarranted trust in AI-predicted answers more than in their own independent decisions without sufficient evidence to justify their trust in the AI answer. When tasks are complex and AI is available to support human privacy reviewers, the reviewers may also overestimate the capabilities of AI and assume it's consistently correct. [71] Human privacy reviewers with access to predicted answers to evaluation questions are likely to trust and agree with the answers because they appear intuitively persuasive, but the reviewers may be more likely to select incorrect answers when polices use contradictory or vague language. [72]

[70] Ren, C., Pardos, Z., & Li, Z. (2024). Human-AI collaboration increases skill tagging speed but degrades accuracy. https://arxiv.org/abs/2403.02259v1.

[71] Kristensen-McLachlan, R., Canavan, M., Kardos, M., Jacobsen, M., & Aarøe, L., Chatbots are not reliable text annotators, https://arxiv.org/abs/2311.05769.

[72] Weisz, Muller, M., Ross, S., Martinez, F., Houde, S., Agarwal, M., Talamadupula, K., & Richards, J. Better together? An evaluation of AI-supported code translation, p. 23, https://arxiv.org/abs/2202.07682.

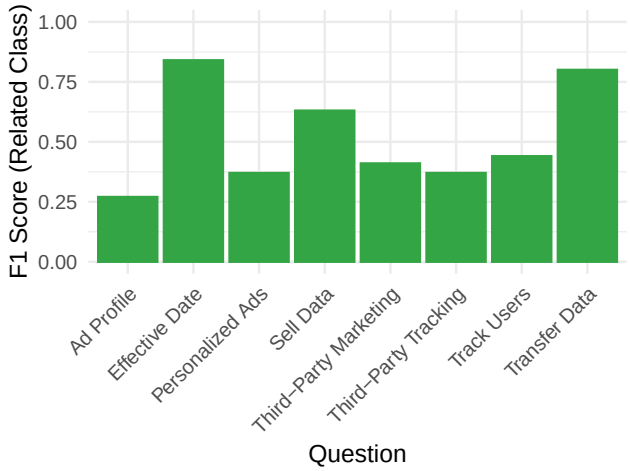
Our initial AI-augmented workflow set out to validate our methodology, as well as identify any potential barriers to increasing the number of questions that additional AI models could cover. As part of this process, we intentionally focused on our core seven privacy rating questions, which cover a wide range of topics but often contain overlapping and nuanced details that we wanted to isolate so that we could ensure our models were identifying the relevant details. We also focused on increasing efficiency of the evaluation process.

In this initial consideration, we chose to include an eighth question—our "Transfer Data" evaluation question, which specifically covers the topic of selling or transferring data as part of a merger, acquisition, or other type of business operational transfer, and often discusses topics like selling personal information as part of either of the aforementioned practices.[73] This question was included for two major reasons. First, it has an incredibly high transparency rate[74] and would present the most accessible opportunity for creating a successful model. Second, it has language that often sounds like it relates to the direct sale of personal information, which our "Data Sold" question covers, and we wanted to validate that both the model for the "Transfer Data" and "Data Sold" questions were appropriately including or excluding the appropriate policy text.

Of course our models will not perform perfectly, but we need to understand the opportunities and challenges the models may present when they're used in our evaluation process. As such, we made sure to capture operational details. We improved our QA pipeline and review process, and we noted details from our reviewers' experiences to learn where the models are performing well or where they need improvement. This continuous feedback loop and information-gathering process across a diverse set of examples over time was critically important in helping us identify areas of strength and weakness in our AI models. The feedback loop also helps indicate where the Program needs to allocate additional resources for data cleanup and fine-tuning so we can achieve desired F1 model scores for every Basic evaluation question.

It is critical, when evaluating and testing our AI models for correctness, that human privacy experts are involved at each step of the process to identify whether the models annotate the correct or incorrect details to answer each evaluation question. Initially this can look like duplicated effort, but it's actually an important step in validating both the reviewers' accuracy and the models' usefulness.

Figure 1: Initial deployed Q2 2020 model F1 scores for the related class only. Note that the macro average that is typically reported would boost many scores up to 30 points, but would misrepresent model performance. As such we typically choose to report and consider the related class F1 scores as that provides the most useful insight into our problem space.



We also explore model failure modes so we can provide adequate UI and UX features to help guide our reviewers in learning to use these tools effectively and understanding their limitations. This is important because we don't want our human reviewers to become too reliant on a useful but imperfect system. Privacy subject matter experts have the necessary experience and training to interpret privacy policy language and what it means in practice. In addition, experts are able to determine whether our AI models are correctly identifying relevant or irrelevant portions of policy text. Therefore, we rely on our experts and their domain-specific knowledge to refine the AI model, assess the model's behavior, inform the creation of the corpus and alignment process.

As part of our initial pre-deployment and testing, we identified several shortcomings in how our existing segmentation process performed when presented with real-world privacy policies. In many situations, large and sprawling sections of text contained unclear sentence structure, complex conditional clauses, and other issues that prevented our models from giving high-quality predictions.

As a result, we prioritized improved segmentation performance in hopes of improving the models' prediction quality and respective annotation scope. Our interns primarily focused on improving segmenter performance, especially in restructuring the complex nested clauses that show up all too frequently in privacy policies. Now, our improved segmenter can usually detect these complex clauses and reconstruct them into individual sentences, making them easier to understand by both people and machines. Please see the "Custom Segmenter" *Appendix* section for more details on our segmentation process.

[73] Kelly, G., Graham, J., & Garton, S. (2023). Privacy Program Evaluation Framework. Common Sense Media, pp. 100 ("Transfer Data").

[74] *See* Common Sense Privacy Evaluation Framework, https://privacy.commonsense.org/content/resource/publications/2023-privacy-program-evaluation-framework.pdf#section*.143 (observed 88% transparency).

Table 4: Manual and AI-enabled question break-down for a Quick Evaluation

Evaluation Type	Policy Time	Question	Transparency	Qualitative	Annotations	Time per Question	Expected Completion Time
Quick (7) Manual	30 mins	30 secs	30 secs	60 secs	2 mins	4-6 mins	60 mins
Quick (7) AI	0-1 min	30 secs	30 secs	30 secs	30 secs	2-3 mins	15 mins

Additionally, we added abbreviations common in statutory language and privacy policies to a custom tool based on PunktSentenceTokenizer[75], which we use to break policies down into smaller segments. The segmentation issues were not barriers to using the models in production, but it was clear that improved segmenter performance would improve the quality of existing and future model performance.

Artificial Intelligence Deployment

These models were developed and launched in 2020, when the market for AI platforms was not quite as robust as it is today. Considerations about hosting costs and cost predictability led us to develop a worker-based batch queuing system to keep costs predictable and affordable. We built out a system based on a Celery[76] worker queue with a Flask[77] based API for managing queue-related details. This queue management system has a worker pool populated with one worker process hosted on an AWS EC2 GPU instance.[78] With the exception of the EC2 GPU instance, the other technology could be easily hosted on our existing AWS infrastructure with minimal additional overhead. The GPU instance has an estimated additional cost of $144 monthly as of May 2023.

The total additional infrastructure cost was rather modest, and was feasible due to our batch-based processing, which didn't require any real-time response because there were no direct interactions with users. To explain further, our existing process of crawling new policies was modified to submit a job to our AI processing queue, which would segment and attach respective predictions to those policy segments and then store the results attached to the policy crawl. This allowed for AI predictions to be run once per new policy crawl. They would then be available for every downstream user of our incoming policy data. Because we don't need real-time AI predictions, the infrastructure spend was much simpler and more predictable.

Having the data attached to the policies also allowed for a flexible rollout of features as we allowed the AI predictions to run in the background and be attached to all new policy crawls. As a lower priority, we also ran AI predictions on historical policy data. While this automated process was running and backfilling our catalog of policy data with AI suggestions, we built out and tested several different UI and UX treatments for integrating the AI suggestions into our existing policy annotation platform. We integrated our AI predictions into our existing annotation platform and reminded our reviewers that the model's output is just a suggestion—the reviewer is the expert and should remain critical of the model output, which is not a perfect system and makes mistakes.

We're also aware of the research showing that overall system performance and accuracy suffers when humans become too reliant on, or too trusting of, AI systems. As such, we made several design decisions to intentionally add friction and slow the process down to minimize the reviewers becoming overly confident on the AI model output. We display the AI suggestions using color-coded notation to mark model prediction relevancy thresholds: green indicates above 50%, yellow indicates 30–50%, gray indicates 10–30%, and anything with a prediction less than 10% is not shown to the reviewer. These visual signals help privacy reviewers quickly identify which AI suggestions the model is more confident about, and if the model is performing well, they are more likely to be helpful to answering the evaluation question. This improves reviewers' productivity, as they can more quickly narrow down the most relevant supporting evidence for a given question.

Through the privacy evaluation process, AI-suggested annotations that meet certain criteria, as described above, are shown to the human privacy reviewer with various UI/UX treatments. The reviewer must explicitly confirm the AI suggestions and add them as annotations supporting their response to the question, just like our non-AI-augmented review process. This explicit confirmation of the correctness of the model's predictions helps us create a rich feedback loop. Essentially, any AI suggestion that is not included as an annotation can be flagged as incorrect for later analysis, and can inform the corpus building and fine-tuning process for future AI models.

[75] *See* PunktSentenceTokenizer, https://www.nltk.org/api/nltk.tokenize.PunktSentenceTokenizer.html.

[76] Celery is a task queue management system that helps distribute tasks or jobs across resources. *See also* https://docs.celeryq.dev/en/stable/getting-started/introduction.html.

[77] Flask is a web framework that assists in building web-based applications and APIs. *See also* https://flask.palletsprojects.com/en/3.0.x/#.

[78] GPU instances provide for faster training and model inference than CPU-based hardware. *See also* https://docs.aws.amazon.com/dlami/latest/devguide/gpu.html.

Table 5: Quick Evaluation cost breakdown of entirely manual process versus AI-enabled review.

Evaluation Type	Policy Time	Time per Question	Total Time	Cost per question	Cost per Evaluation
Quick (7) Manual	30 mins	4 mins	60 mins	$4.46	$31.25
Quick (7) AI	0 min	2 mins	15 mins	$1.11	$7.81

Artificial Intelligence Quick Evaluation Process

After integrating AI into our privacy evaluation workflow, we significantly increased our throughput of published Quick privacy evaluations and reduced the amount of time required to complete each evaluation. AI-enabled Quick evaluations can be completed without the reviewer having to read the entire privacy policy beforehand. Reviewers can refer to the policy text for validation, or to confirm the absence of policy text if the AI models provide no relevant suggestions.

Table 5 indicates that using AI has resulted in a fourfold increase in Quick evaluation productivity, realized by increasing question level efficiency and providing significant reductions in the time required by reviewers to read and comprehend policies, as well as finding relevant portions of policy text. Questions backed by AI models that provide relevant recommendations can now be answered in 2–3 minutes, down from 4–6 minutes per question. Through experience, privacy reviewers trained to use our AI process become more familiar with the seven evaluation questions they complete for every Quick evaluation, and therefore may only spend a few seconds selecting the correct question in the policy annotator software before moving on to the next step. This decreases the time they spend reading each question and lets them spend more time on reviewing answers and providing more accurate annotations.

With Artificial Intelligence, it currently takes a human privacy reviewer 15 minutes to complete a seven-question Quick privacy evaluation.

For AI-enabled Quick evaluations, when a privacy policy is transparent for a given question, the process looks like the following and typically takes approximately 2 minutes:

- **Question (30 seconds):** The reviewer must read each privacy evaluation question asked, just like a multiple-choice exam question.
- **Transparency (30 seconds):** The reviewer is presented with potentially relevant AI suggestions as described above. If relevant annotations are provided, the reviewer may answer the transparency question "Yes." If no relevant annotations are presented below the question, the reviewer may answer the transparency question "No," or find supporting evidence that the model may have incorrectly classified as "not relevant."
- **Qualitative (30 seconds):** If the question is marked as transparent in the previous step, the reviewer reviews the supporting AI-provided relevant suggestions and decides to answer the question qualitatively with a "Yes" or "No" answer.
- **Annotations (30 seconds):** Finally, if the reviewer determines that the policy transparently discloses the specified practice, then the reviewer must provide evidence to justify their answer in the policy by selecting the relevant AI suggestions.

As a result of integrating AI into our review process, the time required to answer each evaluation question has decreased by twofold, and the 30 minutes spent reading the privacy policy is dramatically reduced for a total approximate fourfold increase in productivity. A reviewer may still need to refer back to the privacy policy text to validate questions without AI suggestions or to ensure the policy is actually non-transparent on the issue. Full-time human expert reviewers are able to complete approximately four Quick evaluations per hour for a cost per evaluation of $7.81 ($31.25 / 4). In order to scale to 10,000 products using AI-enabled Quick evaluations,the cost would be approximately $78,100, with an expected savings of $234,400 compared to manual Quick evaluations without the use of AI.

Integration of artificial intelligence significantly reduces the cost to complete Quick evaluations by 75% and increases rating productivity by fourfold.

Projected Artificial Intelligence Basic Evaluation Process

We plan to build on the experience and validation gained after deploying AI for our Quick evaluation process. We will continue to use AI to streamline our workflow and increase throughput, and we will extend AI-enabled questions to all 28 Basic questions in 2025. While we have

Table 6: Basic evaluation cost breakdown of entirely manual process versus AI-enabled review.

Evaluation Type	Policy Time	Question	Transparency	Qualitative	Annotations	Time per Question	Expected Completion Time
Basic (28) Manual	60 mins	30 secs	1 min	90 secs	3 mins	6-8 mins	240 mins
Basic (28) AI	1 min	30 secs	30 secs	30 secs	30 secs	2-3 mins	60 mins

made tremendous progress on the amount of time it takes to complete a Quick evaluation with AI, we expect to see similar productivity gains per question for Basic evaluations as well. Our goal is to reduce the amount of time spent to complete a Basic evaluation from 4 hours to only 1 hour.

With our improved AI capabilities and the insights gained from our AI-augmented Quick evaluation process, reviewers will be presented with the same UI/UX and AI suggestions that are possibly related to each question. This similarity in the approach to Quick evaluations with AI-enabled questions will reduce the cognitive load and task switching across disparate topics for reviewers. We expect this to increase productivity and lower costs analogous to our efficiency gains realized as part of our AI-augmented Quick evaluation process. This approach is also more aligned with legal comprehension exams such as the Law School Admissions Test (LSAT) Reading Comprehension exam section, where test takers need to spend on average about 1 minute and 30 seconds per question.

We expect that using AI to assist in Basic evaluations will improve productivity and result in the same 75% reduction in cost per question as we saw with our AI-enabled Quick evaluations.

Table 6 indicates that the AI-enabled Basic privacy evaluation is expected to take approximately 1 hour —or about 2 minutes per question—to complete when AI recommendations of relevant annotations are provided for each question. This would result in both a rating (Pass, Warning, Fail) and an overall score to further differentiate products by their various privacy practices. If we can realize these gains, we'd see a fourfold increase in productivity of published Basic evaluations over our current manual approach without AI, and would realize similar cost savings. Each reviewer will continue to earn a fixed rate ($31.25/hour) to read, analyze, and complete a single Basic evaluation. For Basic AI-enabled evaluations, the four-part, multi-step question process for a transparent question is the same as Quick AI-enabled evaluations, but with more questions.

Table 7 shows that the expected cost to review 10,000 products using AI-enabled Basic evaluations would be approximately $312,000, with an expected cost savings of $938,400 compared to Basic evaluations without the use of AI.

Artificial Intelligence-Predicted Answer Models

In addition to creating AI models of recommended relevant policy annotations for each evaluation question, our Program explored AI models that would recommend an answer to each Quick evaluation question. However, we encountered several challenges in creating models that accurately determined the proper qualitative response to a question.

Our existing annotations capture all the substantive details about a particular question, but due to the complex nature of privacy policies and of a company's practices, some annotations may indicate that a given practice *is* engaged in, but other annotations within the same policy may indicate the product *does not* engage in the same practice. Because of the potentially contradictory statements, or statements that may apply in certain contexts or scenarios but not others, the collective answer to all of those practices may indicate either "yes" or "no," but some of the relevant annotations may provide contradictory evidence. As a result, we would need to do an additional pass of corpus annotations to label individual annotations indicating "yes" or "no" qualitative practices to particular contexts.

Additionally, we assume with some evidence that even if we did create models that provide a predicted answer, the amount of time it would take a human reviewer to verify the answer may be the same, or greater, than if the human were to complete the review without using AI at all. Accuracy could also be compromised due to the challenges of human-augmented AI systems.[79] Despite these challenges, we were able to create models that could, in certain narrow circumstances, predict an answer with some level of accuracy for some questions. However, when the models struggled, they were often using irrelevant data to determine an answer or they were making an incorrect prediction that cast doubt on trust in the total system behavior. Based on our experience in this space and our exploration of what a

[79] Ren, C., Pardos, Z., & Li, Z. (2024). Human-AI collaboration increases skill tagging speed but degrades accuracy. https://arxiv.org/abs/2403.02259v1.

Table 7: Projected AI enabled cost for Basic Evaluations.

Evaluation Type	Policy Time	Time per Question	Total Time	Cost per question	Cost per Evaluation
Basic (28) Manual	60 mins	6 mins	180 mins	$4.46	$125
Basic (28) AI	0 min	2 mins	60 mins	$1.11	$31.25

human- and AI-augmented system with models like this could look like, we decided to focus our efforts on refining models that find the relevant annotations, because we already know that this is an effective and time-efficient way to maximize the impact of our limited resources.

Additionally, from our research and experience in assessing privacy policies, we decided that regardless of what models or processes we create, a human should always be in the loop to ensure complete and accurate identification of relevant portions of the policy text. As such, we tried to maximize the value of machine automation combined with human expertise to create a hybrid system that was more efficient and accurate than either the human or machine automation alone.

Building AI models that identify the relevant portions of text enables opportunities for other use cases and acts as a potential pipeline for models that may attempt more advanced tasks, such as fully answering a question. We predict that high-performing models that can identify relevant portions of text will be a valuable entry point to creating future models, which may be able to answer a question accurately or perform more advanced capabilities. This approach will likely make verification of more advanced models by a human more straightforward, since the human likely won't need to consider the entire context of a privacy policy and would instead only need to consider the input provided by the models that have identified the appropriate portions of text. Separating concerns like this makes verification of a properly functioning AI system simpler to inspect and correct at each point.

Artificial Intelligence Feedback Loop

Before we started using AI in our evaluation process, the Program relied on experts manually reviewing the evaluations that had been completed by other privacy experts. This process was rather labor-intensive and tended to focus on specific, high-profile questions, such as those that determine a product's rating.[80] With our integration of AI and the subsequent increase in scale, we turned our attention to refining our Quality Assurance (QA) process.

There were two major reasons we focused on this area. First, we needed to ensure that we had some way to quickly verify that we are adequately justifying our question responses with sufficient annotations confirmed by expert reviewers. Second, we wanted to be sure that our AI feedback loop had an additional expert reviewing the AI-provided suggestions that weren't included. It's important to understand that while AI can improve our Program's efficiency, maintaining these systems includes additional costs over time. As these models are deployed in a changing and dynamic system, they will also need to be refined and updated with new information that reflects how the world is changing in the respective domain or intended use. So it's critical to create a feedback loop where the shared decisions of the human- and AI-augmented system inform future models. Rather than create separate systems and processes, we attempted to integrate a QA feedback loop into our existing workflows.

We improved the Program's existing QA process by creating a new QA reporting tool within our policy annotator software that quickly and easily allows a second human expert privacy reviewer to review all the answers that were completed by another reviewer. The first reviewer's answer is displayed alongside the AI-suggested relevant annotations for each question. We feel this QA method better aligns with the overall goal of increasing productivity, while also maintaining the same or higher level of accuracy for each evaluation question. Instead of displaying a "predicted answer" to a privacy evaluation question, the QA reviewer is shown the first privacy reviewer's "actual answer" to each evaluation question. The QA reviewer then "grades" the first reviewer's answers by quickly verifying that the annotated evidence they provided matches one or more of the AI-suggested relevant annotations for that question.

Table 8 shows that this QA process takes approximately 1–2 minutes per question, given the number of annotations provided per question, and is similar to the process used by human reviewers for AI-enabled Quick evaluations but with time reductions of 50% at each substep thanks to the "grading" or supervised nature of the process. We believe this time reduction is the result of summarization and task decomposition.[81] The QA process limits the reviewer's cognitive load to the specific task of review, which does not involve reading the policy, answering questions, or evaluating the relevancy of

[80] Common Sense Media, Privacy Program, Privacy Ratings. https://privacy.commonsense.org/resource/privacy-ratings.

[81] Wu, J., Ouyang, L., Ziegler, D., Stiennon, N., Lowe, R., Leike, J., & Christiano, P. (2021). Recursively summarizing books with human feedback. https://arxiv.org/pdf/2109.10862.

AI recommendations required in Quick evaluations. The QA reviewer only reviews summaries of the evaluation question (the two-word description), the reviewer's answer (transparency/qualitative), the reviewer's evidence (the annotations), and the AI-recommended annotations all together, resulting in a reasonably sized chunk of information that can be more easily processed.

As a result, we take the top-level evaluation task and deconstruct it into several smaller subtasks, whose answers help the human QA reviewer evaluate the top-level task of whether the evaluation's answers are correct or incorrect and why. We believe this transition from an entirely manual open book-type exam format to a partially automated multiple-choice (AI-enabled) exam, and now a QA-driven graded exam, is the best approach. It maximizes productivity at a high level of accuracy and quality to ensure the correct answer is provided for each evaluation question.

As part of our model training process we built a first-generation model using just the privacy evaluation annotations provided by our privacy reviewers. An additional privacy expert[82] reviews the model predictions relative to the privacy reviewer's annotations and corrects or reconciles any disagreements. In this process, a disagreement between model and reviewer is any situation where the model predicts something that does not align with what our human experts have indicated. The process essentially looks like the following:

1. Build the new model.
2. For each model and each existing evaluation, break each policy into individual segments.
3. For each of those segments, run the model prediction against that segment.
4. For each combined model prediction and segment, assess whether the model prediction aligns with the human inclusion of the segment in the list of relevant annotations.
5. If the model disagrees with the human, either add the missing segment or remove the segment from the list of annotations.

This model building and refinement process is iterative. To date, we have found that after at least eight generations of models and cleanup passes, the model performance is refined to a point where it is predictable and where we can provide an adequate assessment of how the model may perform within our evaluation process, or if it needs more work.

To further explain how we determine a model and reviewer disagreement, we set a relevancy threshold—typically 20%, but it may vary depending on model iteration and the specific evaluation question we are building a model for. Then, if the model prediction for a question is over or under this threshold as described below, we consider it an annotation and prediction disagreement. We are more concerned about missing information, so we use the 20% threshold slightly differently depending on context. If the human has not provided an annotation, and the model thinks it is relevant with a prediction over 20%, then we have a third expert review the segment and consider including it as a relevant annotation. This essentially creates a window of 80% relevancy, where an additional human reviews potentially missed segments by the first human reviewer. On the other hand, if the privacy reviewer has provided an annotation and the model predicts that it is relevant with a prediction under 20%, we show this annotation to the third expert so they can consider omitting the annotation from the list of supporting evidence. This way, the privacy expert only needs to review situations where the model prediction and the privacy reviewer disagree considerably and in ways that would be detrimental to our intended use of these models.

An added effect of this process is that the privacy expert who is cleaning up and aligning our annotations relative to model performance can start to understand the types of scenarios that the models struggle with, and can make informed decisions to attempt to train out that behavior in future models or to document areas where we know the model does not perform well. Please see the *Appendix* for more details on F1 scores and annotation cleanup and alignment.

Artificial Intelligence Cost Analysis

The Privacy Program's research and development costs to design, build, test, and deploy our AI models were modest given our small team size and limited budget as a nonprofit, especially when compared with dedicated multi-person teams at for-profit companies with large research and development budgets. Our limited expenditure to build and deploy our AI models was largely driven by our grant-funded budget. This imposed time and resource constraints on our program because we are only one of many groups within Common Sense Media, the larger nonprofit organization. These constraints required creative problem solving and the use of free open-source software and AI solutions with our small yet mighty team of highly skilled employees. Our Program was able to develop and deploy AI successfully because of the unique intersectional domain expertise of the Program's staff in the fields of artificial intelligence, computer science, law, education, and privacy.

The following analysis approximates the organization's total initial investment in AI development costs. However, annual recurring program and administrative costs, including the salaries of the Program's full-time staff and the tech stack used to operationalize the production of privacy evaluations and train privacy reviewers ("Contractors"), are not included in our AI cost-benefit analysis. Our Privacy Program staff, privacy tech stack

[82] Typically this is a third person, but in a handful of evaluations and annotation data this may only be a second expert.

Table 8: Quality Assurance (QA) AI-enabled question break-down for a Quick Evaluation.

Type	Policy Time	Question	Transparency	Qualitative	Annotations	Time per Question
QA Quick (AI-enabled)	0 mins	15 secs	15 secs	15 secs	15 secs	1-2 mins

hosting, product validation, and independent contractor training and evaluation costs are recurring annual costs that are expected to occur regardless of whether AI is integrated into our privacy evaluation process. In addition, the Program pays an operational cost to maintain published privacy evaluations and update them when privacy policies change. We expect to realize the gains from our AI-assisted evaluation process in scaling to evaluate more products with more depth, and in maintaining the approximate 60% of our published evaluations that require annual updates.

The Privacy Program's AI personnel costs were primarily structured around the work of one full-time Principal Engineer staff member, who developed our AI models and data infrastructure with help from two seasonal interns. The Principal Engineer was responsible for managing the entire AI research and development pipeline. This engineer also served as the Privacy Program's artificial intelligence expert and as one of the Program's privacy subject matter experts. The Principal Engineer completed the AI research and development in a part-time capacity alongside their other responsibilities at Common Sense Media, which included several projects within the Privacy Program and across teams in the organization. In total, one Principal Engineer and one Development Operations (DevOps) Engineer were allocated to this project.

Both of these engineers have other responsibilities within the organization, so the total amount of time they spent on AI work is certainly an approximation. While having such a small team enables some efficiency due to simpler channels of communication and fewer coordination efforts, we want to note that this is a considerable amount of work, and because of the limited resources, progress and advancements can sometimes roll out more slowly than they would coming from better-resourced programs and projects.

The Principal Engineer designed, deployed, fine-tuned, and tested the AI models over several months before integrating the models inside the policy annotation software. This allowed for further testing by human privacy reviewers as part of their privacy evaluation workflow. The Principal Engineer was also responsible for the supervision of two summer interns in 2019 and 2022. Both interns performed some initial research and a survey of the AI landscape and made valuable contributions to our custom segmenter. [83]

[83]Thank you to our 2019 intern, Benjamin Fleischmann, and our 2022 intern, Olivia Figueira, who both made valuable contributions.

Because the Privacy Program would have completed privacy evaluations regardless of the inclusion of AI, and those same evaluations would contain annotations to provide an audit trail, we did not calculate what it would specifically cost to develop our AI training corpus. That said, the initial training corpus consisted of 629 evaluations, and has since been expanded to include 1,327 privacy evaluations at an estimated cost of $200,000. The corpus-building pipeline is fairly stable at this point, so the majority of development costs, along with the contributions by our interns, are already incorporated, and we can iterate on our AI without too much consideration for additional development costs.

The hardware costs to train our AI models were limited to local offline training using consumer-grade personal computer hardware with discrete GPUs. It cost a fixed amount to integrate our AI models into our existing privacy policy annotation software and respective APIs, and that cost included the technology infrastructure price per month to deploy and host the AI models. The hardware used to fine-tune the AI models consisted of dual GeForce RTX 2080 Ti GPUs (graphics processing units) purchased in 2020 and running on a Linux-based personal computer workstation with an approximate value of $2,500. The AI model's output was integrated into the Program's existing policy-crawling process, and as a result is available downstream in our annotator software workflow. The additional infrastructure needed to host the AI pipeline is a modest $144 per month and includes access to GPU resources. The other AI-related features are bundled with our existing hosting infrastructure, whose price is approximately $520 per month and is a fixed cost that would be necessary regardless of whether we integrated AI into our workflow.

Evaluation Scale & Sustainability

As previously discussed, our Quick evaluation manual cost per evaluation is approximately $31.25, and after AI was integrated into the process for the entire set of questions covering the Quick evaluation, the cost went down to approximately $7.81. Along with this savings, we also achieved a few additional gains in increased scale and opportunity to include more experts in each review process. Before we introduced AI into the process, performing reviews on thousands of products put significant pressures on our publishing workflow, limiting opportunities for additional experts to review evaluations prior to publishing. With our increased efficiencies, we have made

concerted efforts to refine and improve our QA process to ensure that all published evaluations maintain our expected level of accuracy and completeness, and continue to include relevant annotations that provide the evidence for why we answered each evaluation question the way we did.

With the proven success of integrating AI into the smaller seven-question Quick evaluation, we are now hoping to realize similar gains by expanding our set of AI-covered questions to include those in our 28-question Basic evaluation. Increasing the number of products evaluated with the more nuanced and complete 28-question evaluation will allow us to make more deeply informed decisions about products. The Quick evaluation is sufficient to create a Pass, Warning, or Fail privacy rating, but offers very little additional insight into the broad range of practices needed to make more meaningful and informed privacy decisions. Questions in the Basic evaluation include much more nuance and detail, which is necessary when making decisions that can impact children's health, well-being, and healthy development. This level of basic information is similar to the amount of detail required for nutrition labels, and is critical for people making decisions on behalf of children and students, as is the case for parents and educators.

Maintaining an up-to-date and accurate reflection of a significant number of products' current privacy practices that reflect their current policies is a considerable challenge. Doing so sustainably at a sufficient level of detail and accuracy presents further challenges. The Privacy Program's methodical and detailed approach has been shown to scale well for our seven-question Quick evaluation with the inclusion of AI. Expanding on this success, we plan to incorporate a similar approach into our Basic evaluation process. As discussed above in the "*Projected Artificial Intelligence Basic Evaluation*" section, we are hoping to achieve a similar 75% reduction in the cost of publishing Basic privacy evaluations. This should help us achieve two major goals: to provide more detailed privacy evaluations for more products, and to maintain those evaluations to reflect current practices as company's privacy policies change over time.

Conclusion

The Privacy Program's use of AI has created a more sustainable path to evaluating privacy policies over time in tandem with human privacy reviewers. We successfully increased productivity by using AI to support and augment human privacy reviewers in more effectively rating product privacy policies at scale.

Our organizational goal is to publish privacy evaluations for 10,000 products, which means that all products reviewed by Common Sense would include a privacy rating. Without AI assistance, this would cost the Privacy Program approximately $312,000 to scale to 10,000 Quick evaluations, and approximately $1,250,000 for the same number of Basic evaluations. After integrating AI, our new cost to review 10,000 products using AI-enabled Quick evaluations will be approximately $78,100, with an expected savings of $234,400. Additional savings would be expected with annual maintenance costs, because the Program would need to update evaluations when policies change. Our estimated cost to complete AI-assisted Basic evaluations of 10,000 products would be approximately $312,000, with an expected cost savings of $938,400 compared to doing the work without AI assistance.

The gains in both scale and efficiency from integrating AI into our Quick evaluation process have allowed us to maintain and improve accuracy by refining our QA process to allow for more human experts to review each privacy evaluation prior to publishing. This hybrid human and artificial intelligence ("Human-AI") approach reduced our cost per product to evaluate different types of privacy evaluations, while simultaneously using supervised machine learning to continuously capture high-quality training data from the annotated privacy policies of each product that was evaluated. Our modest investment in research, development, and integration of AI has proven to increase our productivity. Our existing success will certainly guide our expansion to AI-enabled Basic privacy evaluations, but we also hope that our model of using intentional, non-generative AI solutions can help other researchers, practitioners, experts, or regulators to achieve measurable and evidence-based efficiencies of scale. Our approach has shown it is possible to use human privacy reviewers in combination with AI to increase the number of privacy policies we can evaluate without sacrificing accuracy.

Appendix

Custom Segmenter

The NLTK project[84] has a module for breaking text into sentence-level detail called Punkt Sentence Tokenizer.[85] However, it was trained on data[86] that is not representative of all the abbreviations we encounter in privacy policies. As such, we needed to amend the list of known abbreviations to better segment text that we see in privacy policies. This includes common abbreviations such as i.e. or e.g., but also abbreviations that frequently show up as statute references (for example cal., civ., c.f.r., u.s.c., u.s.c.a., u.s.c.s., stat., and more.).

Additionally, privacy policies frequently use complex lists with hanging clauses that are difficult for humans to understand and present challenges for capturing the appropriate annotation context. As an egregious example, structures like the following often show up in privacy policies:

```
You must not:
    * reverse engineer, de-compile, hack, disable, disrupt,
      interfere with, disassemble, copy, decrypt,
      reassemble, supplement, translate, adapt or
      enhance any of the PCI Property or the Services;
    * create a link, character name or label, or otherwise
      upload to or transmit from Acme or the Services
      any content, link or anything else that (if
      reproduced, published, transmitted or used) may:
        * be defamatory, threatening, abusive, harassing,
          hateful, obscene, pornographic, harmful or
          invasive of anyone's privacy, or excessively
          violent,
        * violate any law including intellectual property,
          privacy or other laws;
        * impersonate any person;
        * give rise to civil or other liability; or
        * relate to illegal drugs, weapons, gambling or
          other illegal activities;
    * place any mature content in the "just for kids" or "
      general audience" sections of Acme;
    * upload to or transmit from Acme or the Services any
      data, file, software or link that contains or
      redirects to a virus, Trojan horse, worm or other
      harmful component;
    * use Acme or the Services to do or attempt to do any
      of the following without PCI's prior written
      permission:
        * send spam or other bulk messages;
        * gain unauthorized access to any data, network or
          system;
```

This example is a truncated excerpt presented as one sentence, but is an incredibly complex compound statement that could be better expressed as 25 individual sentences, only 10 of which are shown for brevity's sake. Our custom segmenter detects this "hanging" or compound sentence structure and sub-segments it into individual sentences. This yields a refined suggestion that includes all the appropriate context of the sentence. Our segmenter sub-segments the preceding chunk of text into the following sentences.

1. *You must not: reverse engineer, de-compile, hack, disable, disrupt, interfere with, disassemble, copy, decrypt, reassemble, supplement, translate, adapt or enhance any of the Acme Property or the Services.*
2. *You must not: create a link, character name or label, or otherwise upload to or transmit from Acme or the Services any content, link or anything else that (if reproduced, published, transmitted or used) may: be defamatory, threatening, abusive, harassing, hateful, obscene, pornographic, harmful or invasive of anyone's privacy, or excessively violent,*
3. *You must not: create a link, character name or label, or otherwise upload to or transmit from Acme or the Services any content, link or anything else that (if reproduced, published, transmitted or used) may: violate any law including intellectual property, privacy or other laws.*
4. *You must not: create a link, character name or label, or otherwise upload to or transmit from Acme or the Services any content, link or anything else that (if reproduced, published, transmitted or used) may: impersonate any person.*
5. *You must not: create a link, character name or label, or otherwise upload to or transmit from Acme or the Services any content, link or anything else that (if reproduced, published, transmitted or used) may: give rise to civil or other liability.*
6. *You must not: create a link, character name or label, or otherwise upload to or transmit from Acme or the Services any content, link or anything else that (if reproduced, published, transmitted or used) may: relate to illegal drugs, weapons, gambling or other illegal activities.*
7. *You must not: place any mature content in the "just for kids" or "general audience" sections of Acme.*
8. *You must not: upload to or transmit from Acme or the Services any data, file, software or link that contains or redirects to a virus, Trojan horse, worm or other harmful component.*
9. *You must not: use Acme or the Services to do or attempt to do any of the following without Acme's prior written permission: send spam or other bulk messages.*

[84] Natural Language Toolkit, https://www.nltk.org.

[85] Punkt Sentence Tokenizer, https://www.nltk.org/api/nltk.tokenize.punkt.html?highlight=punkt#module-nltk.tokenize.punkt.

[86] NLTK Sentence tokenizer does not tokenize properly if there exists "e.g." or "i.e." in the sentence. #2543, https://github.com/nltk/nltk/issues/2543.

10. *You must not: use Acme or the Services to do or attempt to do any of the following without Acme's prior written permission: gain unauthorized access to any data, network or system.*

Our custom segmenter provides the entire context of an individual sentence, which helps with proper annotation of just the relevant portions of text, and can contribute important information to our data pipeline so that both machines and humans can make more informed and accurate decisions.

F1 Scores

When discussing AI model performance, especially when data is heavily skewed, it is important to use metrics that provide more insight than a simple measure like accuracy. The models we are attempting to build identify portions of text related to a specific question. In our absolute best-case scenario, around 4% of a policy's text will be relevant to a given question, but the amount of relevant text is typically closer to 1% or less. If our model predicted that every portion of a policy was not related to a given question, but 4% of the policy was actually related to the question, that model would be 96% accurate but essentially provide no meaningful insight into our problem. It would be a highly accurate but generally useless model. As such, we need a method better than accuracy to discuss model performance, especially in a lab environment. In machine learning work, the F1 score is frequently used to assess model performance.

To explain what an F1 score is, we will first need to talk about two other metrics: precision and recall. *Precision* is the number of truly relevant policy annotations divided by the number of all policy segments the model indicated were relevant, including segments where the model prediction incorrectly predicts a segment is related. *Recall* is the number of truly relevant policy segments divided by the number of all segments the model predicts as relevant. For our purposes, a high recall value is more valuable in making sure we do not miss any relevant policy text. Of course, this needs to be balanced by precision, since we do not want our human reviewers to have to review large amounts of irrelevant text. The F1 score is a way to combine precision and recall into a single value that provides insight into how a model might behave in other scenarios.

While F1 scores can be useful for discussing theoretical performance, they have limitations when considering real-world scenarios and the complex data we may encounter. Additionally, in isolation, F1 scores do not necessarily reflect total system performance—even models with low F1 scores may be able to be effectively integrated with other models, system designs, and UI or UX treatments. This can lead to a better overall system performance than what a naive F1 score assessment may indicate.

We choose to report only the "related class" F1 score, because in our use case we are most concerned about properly identifying relevant portions of a policy text, and as such do not need to focus on the "macro average" F1 score. In nearly all other cases, the "not related" F1 score is .99 or 1.0, which provides little insight into how our models may behave in the real world. The following chart is shared for illustrative purposes to indicate how our F1 scores may compare to other researchers' F1 scores.

Beyond F1 Scores

Calculating an F1 score assumes a naive threshold of 50%, where anything under 50% indicates not related and any prediction over 50% indicates related. But depending on how the model actually performs, we may want to arbitrarily draw a boundary at a different point than 50%. As part of our model-building process, we calculate histograms that show a brief visual summary of how a model performs relative to our human-provided annotations. The two charts give us a visual representation of how the model performs on our test data, and provides a much richer understanding than an F1 score in isolation does.

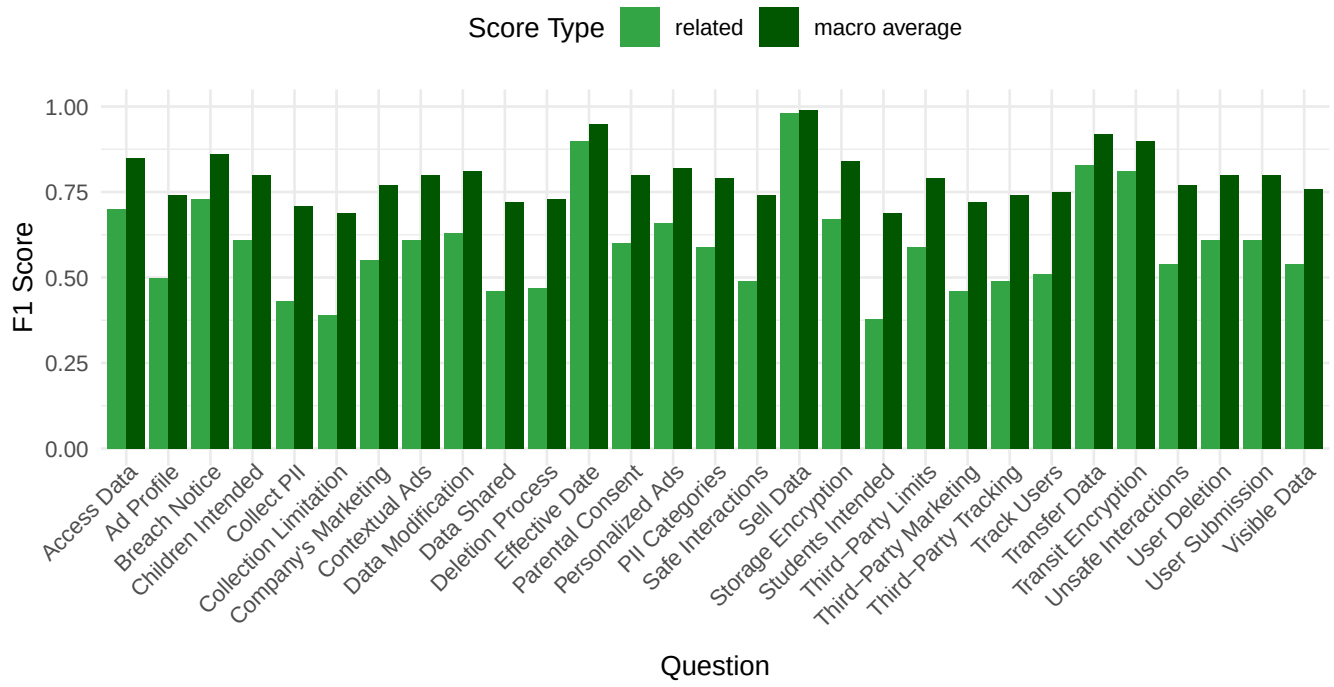
The first example is our "Sell Data" model, which has had the most refinement and has undergone several passes of annotation and segmentation data alignment. The F1 scores indicate very good performance and the two charts show us the model prediction probability spreads for both the "notRelated" and the "Related" class. As we can see from both charts, which have a large bar representing the vast majority of data far to the right, this model is very accurate for both classes and very confident about the predictions. The language required to discuss the practice of "selling data" is one of the most regulated in the industry, while what we see in practice includes obfuscating and confusing language. Despite this complexity, there is, generally speaking, a language style that is fairly consistent across the industry. We also see that we have 1,513 examples of what discussing selling data looks like in our "Sell Data" test or validation data set.

Table 9: Sell Data model performance assessment showing precision, recall, and F1-score for both "notRelated" and "related" classes of data as well as combination metrics for both classes.

	precision	recall	f1	support
notRelated	1.00	1.00	1.00	35525
related	0.99	0.97	0.98	1513
macro avg	1.00	0.99	0.99	37038
weighted avg	1.00	1.00	1.00	37038

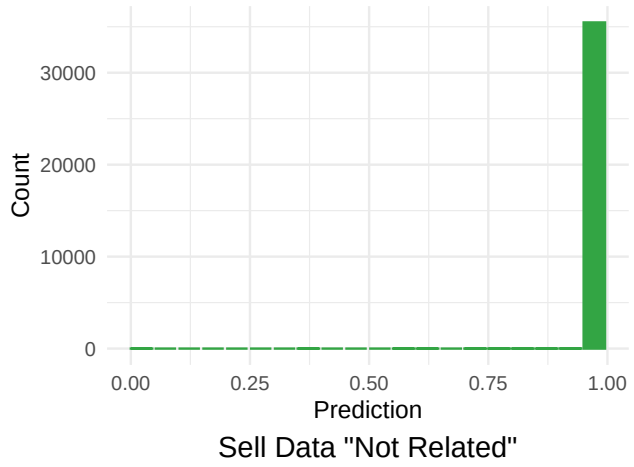
In figure 3, each bar indicates the number of predictions falling into the prediction range as indicated on the x-

Figure 2: Q2 2024 F1 Scores for basic questions. This is a draft placeholder data snapshot pending updated 2024 numbers expected Q4 2024. In this placeholder figure only the Sell Data model has received cleanup/alignment passes.



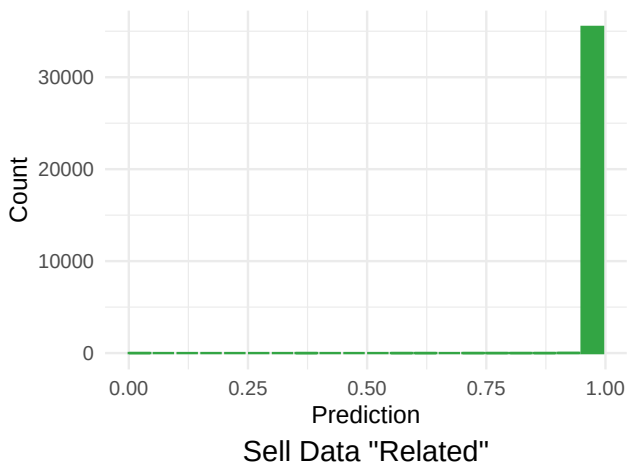
axis. The large bar completely to the right indicates that nearly all of the sample text is appropriately predicted to be "not related" with a high prediction likelihood.

Figure 3: Histogram chart of "Sell Data" model prediction probabilities for text indicated as "not related" by human provide annotations.



In figure 4, each bar indicates the number of predictions falling into the prediction range as indicated on the x-axis. The large bar completely to the right indicates that nearly all of the sample text is appropriately predicted to be not related with a high prediction likelihood. Note that there is a very small sample of predictions across all ranges, meaning the model has some uncertainty. We see a more notable uptick in predictions near the 0.0 mark, indicating that the model is telling us that some human-provided annotations were marked as related, but the model is predicting that they are likely unrelated.

Figure 4: Histogram chart of "Sell Data" model prediction probabilities for text indicated as "related" by human provided annotations.



The second example is of a model that does not perform quite as well as our "Sell Data" model. This model is an unrefined model for "Third-Party Limits."

Initially we see two major differences in the number of total samples in our validation set. For one, we have about 71% as many total examples (25226/35525) of the the "Sell Data" model, and roughly 27% (406/1513) as

many examples of what "Third-Party Limits" policy text looks like as compared to what "Sell Data" policy text looks like. It should be noted that our test or validation sets reflect the same proportions of related vs. notRelated text we see in our training data sets for each respective question. So, we shouldn't be too surprised to see that our "Third-Party Limits" model does not perform quite as well as the "Sell Data" model on the "related" class, since we have considerably fewer examples to learn from.

Secondly, there are not as many statutes and regulations to indicate the requirements companies must follow when disclosing their practices with respect to the practice of "Third-Party Limits." Interestingly, the "notRelated" class performance is comparable for both models. This disproportionate amount of data results in a macro or weighted average F1 score being skewed and inflating the perceived performance of our model by 20 to 40 points. This is a good example of why we typically only report the "related" class performance, as it gives us a better understanding of how a model may perform on the scenarios we are most concerned about getting correct.

The two histogram charts below provide a lot more information about the ways this model is underperforming. Unfortunately, our figure 6 shows a large bar far to the left, indicating that the model is confidently incorrect and predicts a large amount of "related" text is likely "not related." We can also see a considerable spike in data on the right side. This is promising and implies that with several data passes and annotation and segment alignment, this model may be able to achieve performance more like our "Sell Data" question. Our annotation and refinement pipeline only presents the human reviewer with data that the model is confidently incorrect on. In our graphs, that is all data to the left of those first 0.2 marks on the x-axis. This helps us maximize our human expert time by focusing only on those issues where the model predictions and human-provided annotations disagree strongly. Through several iterations of refining the annotation data and rebuilding models, we work our way toward better model performance, focusing on those details where the disagreement is strongest.

As discussed above in the sections "Road to Production Ready" and "Artificial Intelligence Feedback Loop," there are various reasons why the segment prediction and related annotations may disagree. To address this, we bring in another privacy expert to align the mismatches and reconcile disagreements.

As the model performance improves and becomes more refined, we sometimes see instances where the model does not have enough examples of some details to adequately learn. For example, our "Data Sold" model sometimes incorrectly predicts some statements around "users not being allowed to sell their accounts" or "users selling products" as being related to the "Data Sold" question. We have not attempted to refine this process, largely because we do not have enough examples of how this information is discussed. Further, the number of products that discuss these details and the number of annotations that humans may have to potentially reject is relatively small compared to the value the model otherwise provides.

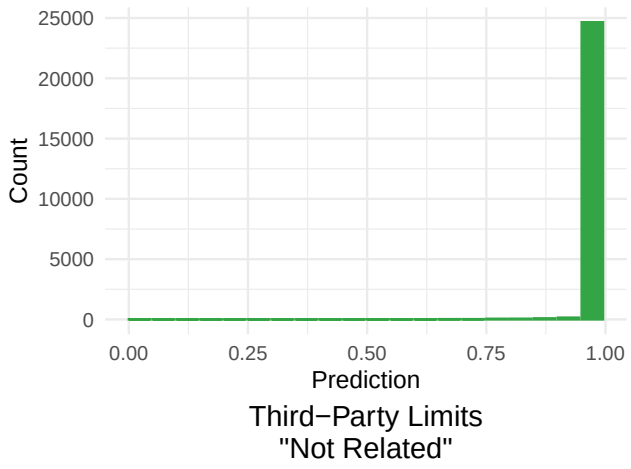
Rather than attempting to oversample, or otherwise mitigate the few examples of rare text, we have decided to document this behavior and intend to provide additional documentation or guidance on how to use the respective models effectively as part of the evaluation process. We will likely revisit this approach, but as of now any mitigation efforts we have attempted have degraded model performance in ways that created more work and a final model that performed worse. In the "Annotation Cleanup & Alignment" section, we discuss why in some contexts and for some topics it may be better for a model *not* to be confident about whether text is related.

Table 10: Third-Party Limits model performance assessment showing precision, recall, and f1-score for both "notRelated" and "related" classes of data as well as combination metrics for both classes.

	precision	recall	f1	support
notRelated	0.99	1.00	0.99	25226
related	0.69	0.51	0.59	406
macro avg	0.84	0.76	0.79	25632
weighted avg	0.99	0.99	0.99	25632

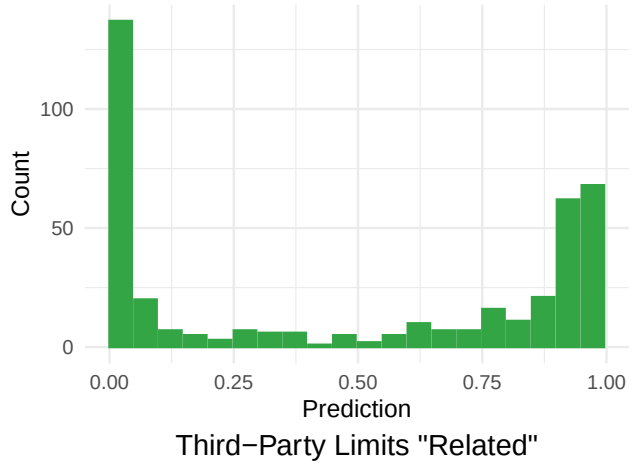
In figure 5, each bar indicates the amount of predictions falling in the prediction range as shown on the x-axis. The large bar completely to the right tells us that nearly all of the sample text is appropriately predicted to be not related with a high prediction likelihood. As compared to the corresponding "Data Sold" chart, we can see a longer trailing edge of uncertainty,which is likely difficult to see on this scale, starting near the 0.75 mark.

Figure 5: Histogram chart of "Third-Party Limits" model prediction probabilities for text indicated as "not related" by human provided annotations.



In figure 6, each bar shows the number of predictions falling in the prediction range as indicated on the x-axis. The large bar completely to the left tells us that a large number of samples in our validation data are being confidently and incorrectly labeled as "notRelated" even though human reviewers have indicated they are related. As we gain more insight into what types of examples fall into this category, we can refine the annotation cleanup and alignment process and come to a better understanding of the failure modes of this model. Additionally, we see a large spread of predictions across all prediction percentages, with a notable increase in prediction density in the area near 0.75–1.0.

Figure 6: Histogram chart of "Third-Party Limits" model prediction probabilities for text indicated as "related" by human provided annotations.



Annotation Cleanup & Alignment

As part of our corpus building and segmentation process and as part of refining model improvement, most of what we focus on to increase model performance is aligning our provided annotations that are part of our AI corpus with the desired behavior of the model. To do this, the person building the AI model reviews annotations where a model and a human-provided annotation do not agree, or disagree beyond some threshold of AI prediction. As an example, we explore what that looks like for one instance where the AI prediction does not align with the human-provided annotation.

The phrase under consideration is the following:

> *"Please see our 'Children's Privacy Policy,' below, for information about our collection and use of information from children under 16.)"*

The AI model is used with our custom annotation alignment software, and the operator is presented with situations where a human-provided annotation does not align with an AI-suggested segment. For example, if the model makes a prediction that the above segment is 0.64 "not related" to the "Children Intended" question and our human reviewer did not mark the annotation as related to "Children Intended," the operator is presented with the annotation for review, because we set our model threshold for alignment and review below 0.8, and 0.64 is less than 0.8. While technically speaking this segment is not related to our "Children Intended" question, as our question is focused on an under-13 threshold, the operator may make a judgment call and include this annotation to create a model with the desired behavior in the real world. To further explain, when presented with other text that appears to be similar to this particular phrase, there may be ambiguity that should be presented as potentially supporting evidence in some cases. That is, even though this particular example may not be related to our "Children Intended" question, we may want the model to *not* be confident that the annotation is "not related" to "Children Intended."

Having a less confident prediction closer to the 50% threshold gives us quite a bit more information and includes some level of uncertainty, rather than having a model that is confidently correct or confidently incorrect, which may not be the most useful model behavior in a general sense. As such, the operator must carefully consider both the immediate context of the question policy text *and* the larger context of future desired model behavior. The operator of our annotation alignment software must decide to either re-annotate this annotation to indicate it is related to our "Children Intended" question, or to leave the text as is and hope that with refinements to annotations elsewhere, the next model iteration will be more confident in a prediction. This example also shows the need for a human expert, as we likely **do** want the model to provide a non-confident prediction since this is technically unrelated to our "Children Intended" question but may be related in some contexts depending on other policy text.