The Role of Online Forums in Developer Understanding of Privacy Law - A Reddit Case Study

Sara Haghighi
University of Maine
Orono, Maine, USA
sara.haghighi@maine.edu

Clark LaChance
University of Maine
Orono, Maine, USA
clark.lachance@maine.edu

Ali Pourghasemi Fatideh
University of Maine
Orono, Maine, USA
ali.pourghasemi@maine.edu

Travis Breaux
Carnegie Mellon University
Pittsburgh, Pensylvania, USA
tdbreaux@andrew.cmu.edu

Sepideh Ghanavati
University of Maine
Orono, Maine, USA
sepideh.ghanavati@maine.edu

Abstract

Software practitioners use online forums to navigate complex and often ambiguous legal privacy requirements, yet little is known about their professional backgrounds, what challenges they face, and how they use and assess the credibility of the advice received, or how they resolve ambiguities in posts. We report the findings of a survey of 223 Reddit users from regulatory-focused subreddits, complemented by a qualitative analysis of 2,248 posts and responses. Our results show that, despite holding privacy-related certifications, most participants frequently use forums to seek legal advice. Key challenges reported or identified include implementing a data protection impact assessment, reporting a data breach, and obtaining cookie consent. Reddit users often assess credibility by reviewing respondents' post history, verifying sources cited, trusting advice from recognized experts, and following up for clarity before responding. We highlight research and educational directions to bridge gaps in support needed for regulatory compliance guidance.

Keywords

GDPR, Reddit, Survey, Privacy Compliance

1 Introduction

Comprehensive privacy regulations, such as the EU General Data Protection Regulation (GDPR) [15], require organizations to adopt accountable privacy-preserving practices, with significant consequences for non-compliance. Translating these regulations into concrete technical and organizational practices, however, is often complex and ambiguous. To navigate this uncertainty, developers rely on a range of strategies, from consulting legal counsel to seeking informal guidance in online communities [48].

Prior research shows that, particularly in smaller companies, developers may turn to platforms such as Reddit or Stack Overflow for privacy-related advice [38, 45, 48, 62]. These studies report on the types of challenges developers encounter and the advice they receive. However, their focus is mainly limited to privacy concepts in app development, e.g., mobile permission requests, sentiment analysis, data collection and sharing practices, privacy-by-design solutions, and overlook broader compliance discussions (e.g., how to handle data subject access requests or cross-border data transfer) that take place in regulatory-focused communities (e.g., `r/gdpr` or `r/europrivacy`). Furthermore, these studies analyze only posts and comments; thus, they do not capture contextual factors such as participants' demographic backgrounds, how these shape their engagement and impact their perceptions and challenges, and how they apply received advice, ensure its credibility, and resolve ambiguities in practice. Hence, it remains unclear how informal interpretations of privacy regulations develop and affect compliance.

We address these gaps by conducting two complementary studies: (1) a quantitative survey of 223 Reddit users who participate in regulatory-focused subreddits; and (2) a qualitative large-scale analysis of 2,248 posts related to privacy and software development from the same subreddits. For the large-scale analysis, we use in-context learning with both open-source and proprietary large language models (LLMs), to identify posts relevant to both privacy and software development and to classify discussions that describe concrete GDPR compliance challenges. To ensure reliability at scale, we combine automated classification with human-in-the-loop validation. We further integrate self-reported behavior and motivations collected through our survey with the posts' analysis to compare perceived challenges with those discussed online, aiming to better understand how practitioners reason about and act on informal guidance. We also explore how factors such as professional roles influence users' motivations, perceptions of challenges, assessment of the credibility of shared advice, and how they resolve ambiguities. Lastly, we demonstrate how credibility factors best predict Reddit users' decision-making behavior (i.e., whether to follow or apply received replies in their work). Our research questions are:

- **RQ1:** What motivates users to engage in privacy and legal discussions on regulatory-focused subreddits, and does this vary based on professional and organizational background?
- **RQ2:** What are the most common challenges users face when seeking compliance or privacy-related advice, and how do they vary by professional and organizational factors?
- **RQ3:** How do users assess the credibility of information shared, resolve ambiguity in posts, and how do these assessments affect their decision-making processes? Does perceived credibility vary based on professional background?

Proceedings on Privacy Enhancing Technologies 2026(4), 321–340

https://doi.org/10.56553/popets-2026-0123

- **RQ4:** How do the challenges reported by Reddit users compare with those identified in the posts?

Our survey results show that most participants are from North America, despite the EU-focused subreddits, and hold privacy-related certifications. We find that user motivations vary by privacy-related background; specifically, those with a privacy-related certification or experience are more motivated to share their knowledge, while others want to seek advice or stay informed *(i.e., RQ1)*. Although most reported having in-house legal counsel, the majority occasionally seek external privacy or legal advice. Survey users reported struggling most with data protection impact assessment (DPIA) and data breach notification, while Reddit post analysis showed consent as the most prominent challenge *(i.e., RQ2 & RQ4)*. We identified six additional challenges in the post analysis that were not reported in the survey, e.g., handling user consent, data processing agreements, and data transfer to a third country *(i.e., RQ4)*. Post analysis showed that responders often referred to guidelines such as EDPB and ICO, while the survey revealed they mostly used the NIST frameworks. These findings underscore the importance of complementary studies in capturing challenges and practices in real-world settings. Users rely on multiple strategies, such as checking the poster's profile and engagement history and cross-checking responses with external sources to assess credibility. However, some users report acting on advice with little or no verification, highlighting potential pathways to non-compliance. This concern is further underscored by contradictory responses identified on Reddit, which require additional resources to verify their accuracy *(i.e., RQ3)*.

Based on our findings, we outline future research and educational directions as well as implications for industry and regulatory bodies.

2 Related Work

Recent studies investigate developer behavior, practices, and experiences in privacy engineering activities through large-scale surveys [2, 39, 48, 50, 56], or targeted interviews [2, 12, 24, 28, 29, 33, 57].

Tahaei et al. [56, 57] and Horstmann et al. [28] examine privacy challenges through interviews and programming studies, identify issues like poor privacy culture and communication barriers as the main challenges, and emphasize the importance of "privacy champions". The presence of a privacy officer and privacy education are shown to increase privacy awareness and the adoption of privacy-enhancing tools [48]. In some cases, developers were found to not think about privacy early in development activities, or to consider it simply a security concept [24, 48]. In a privacy-related programming study with 30 developers, Horstmann et al. [29] found that they rarely consider privacy in their code, and the majority failed to have privacy-compliant solutions. Finally, studies highlight developer privacy concerns [37], challenges in regulatory compliance [2, 59], and that they may prioritize app store requirements over laws, relying on the stores to flag privacy issues [2, 61].

Other works focus on forum analysis to uncover developers' security and privacy concerns. Tahaei et al. [55, 58, 62] explored privacy issues on Stack Overflow, and found that most concerns involve permissions, authentication, and third-party libraries. Their results show that Google and Apple Privacy labels increased the number of privacy questions, particularly regarding consent [62]. Li et al. [38] and Parsons et al. [45] examined `r/androiddev`, `r/iOSProgramming`, and `r/webdev` and found that privacy discussions are infrequent and often motivated by regulatory pressures or app store policies. They focused on trends and sentiments, rather than compliance challenges. Kumar et al. [41] conducted sentiment and temporal analyses of privacy discussions across various mental health subreddits and found that privacy discussions increased significantly over time, and sentiment varied depending on the privacy themes.

Table 1: Comparison of Related Work and Our Study

PS = Platform Studied, US = User Study, MR = Mixed Roles, GM = GDPR Mapping, TRI = Triangulation. Dev = Only Developers

Literature	PS	US	MR	GM	TRI
Li et al. [38]	r/androiddev	×	×	Partial	×
Parsons et al. [45]	Reddit mobile+web	×	×	Partial	×
Santos et al. [49]	r/gdpr (7 posts)	×	×	×	×
Tahaei et al. [55, 58, 62]	Stack Overflow	×	×	Partial	×
Horstmann et al. [28, 29]	Interviews + Survey	✓	Dev	✓	×
Serafini et al. [50]	Survey	✓	Dev	×	×
Kumar et al. [41]	Reddit Mental Health	×	×	Partial	×
Prybylo et al. [48]	Survey	✓	✓	×	×
Pessianzadeh et al. [47]	Reddit discussions	×	✓	×	×
Amini et al. [3]	Reddit discussions	×	×	×	×
This Work	Regulatory-focused subreddits + Survey	✓	✓	✓	✓

Some studies examine trust and credibility in Reddit discussions using large-scale computational approaches, analyzing how public perceptions evolve, how credibility is assessed, and how users interact across different groups [3, 22, 47]. However, these studies focus on general attitudes toward technologies (e.g., AI systems) rather than GDPR-specific compliance challenges, and do not combine survey-based insights with large-scale post analysis.

Prior work primarily analyzes the types of questions on developer-centric platforms, like Stack Overflow, or subreddits, e.g., `r/webdev` or `r/androiddev` [38, 45, 55, 58, 62], focusing on narrow topics (e.g., mobile app permissions, personal data) or conducting sentiment and trend analysis, rather than identifying specific GDPR compliance challenges such as record of processing activities. As shown in Table 1 and Figure 1, we complement these studies by analyzing regulatory-focused subreddits (e.g., `r/gdpr`, `r/europrivacy`) through a dual study. First, we survey 223 users (including developers as well as legal professionals, privacy officers, and end-users) to uncover the types of GDPR concerns raised, their engagement on Reddit, the strategies for assessing credibility in anonymous environments and resolving ambiguities, and their decision-making processes. Second, we conduct a large-scale analysis of posts using an LLM-based approach to identify GDPR challenges in software development. This combination bridges technical and legal perspectives and offers a more comprehensive view of how diverse actors grapple with GDPR compliance in practice. To the best of our knowledge, our work is the first study to directly engage with Reddit users to explore GDPR-related challenges and motivations as articulated by them within their communities.

3 Study Design and Reddit Posts Analysis

We now present the survey design and Reddit post analysis method. Figure 1 shows our study overview.

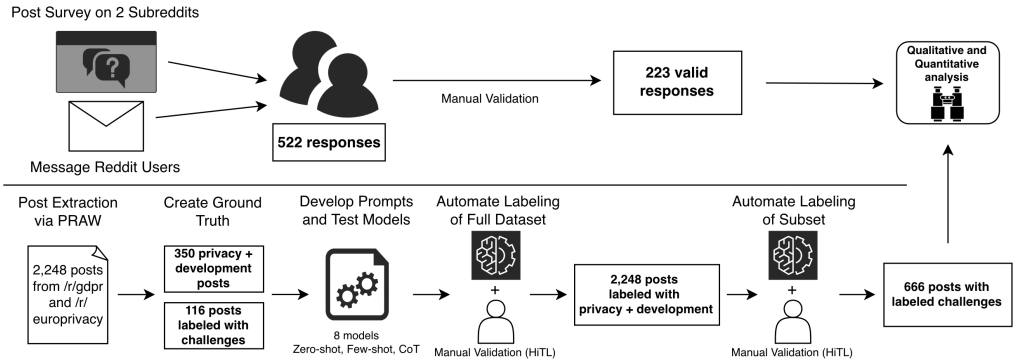


Figure 1: The Overview of Our Study

3.1 Survey Design

The survey was informed by prior research on developer privacy practices [28, 48, 57, 60], how they reason about privacy requirements on forums [38, 45, 49], and by informal discussions with our legal collaborators (i.e., law professors) and a privacy research professional at a major company. We iteratively piloted the survey with members of our university's privacy research group to improve clarity and flow. Based on feedback and pilot response analysis, we refined question wording, clarified terminology, and adjusted Likert-scale options to better align with GDPR compliance constructs identified in prior literature and the CIPL report [8].

The survey was designed with three goals: to reveal the professional backgrounds of users on regulatory-focused subreddits, their motivations, credibility assessments, and decision-making processes, and compliance challenges. We followed the guidelines in [48] to minimize length and focus on users' interactions. Thus, we tailored the questions to user responses. For example, given *Q17: "When interacting on Reddit, what do you do typically? (e.g., creating the original posts, replying to other users' posts, or reading the posts)"*, we present follow-up questions based on the given response. The survey is organized into blocks as follows, using Qualtrics:

- *Professional Background and Interactions:* We ask about participants' professional background (e.g., roles and company location), their privacy certifications, and the frequency and types of their engagements when seeking GDPR advice.
- *Motivation for Interaction:* Aligned with RQ1, we included questions that captured participants' motivations for interacting on Reddit based on their interaction types.
- *Perceptions and Challenges:* To address RQ2, we investigate the key obstacles participants face when implementing GDPR compliance and their perceived barriers to applying externally sourced privacy or legal advice in practice.
- *Credibility Evaluation and Ambiguity Resolutions Strategies:* For RQ3, we examine how users assess the credibility of advice shared and resolve unclear posts, and whether this assessment impacts their decision-making process.

Demographic Question: We placed questions about their demographics *(Q31–Q35)*, including age, gender, education level, degree, etc., at the end of the survey to reduce early dropout and minimize bias in how participants respond to subsequent questions [20].

3.1.1 Data Collection. We recruited participants by posting a flyer on `r/gdpr` and `r/europrivacy` for one week in September 2024, and by directly messaging Reddit users. We selected `r/gdpr` and `r/europrivacy` because these subreddits have 20 times more members than smaller subreddits (e.g., `r/ccpa`) and focus specifically on regulatory compliance topics. This focus contrasts with `r/privacy`, which broadly covers privacy-related topics, and subreddits like `r/androiddev` that focus on device and platform developer challenges, such as how to request permissions. To reach active contributors and increase response rates, we also sent direct messages to selected users. These users were manually identified based on visible contributions (e.g., frequent posting, upvoted responses, and commenting) within `r/gdpr`, `r/europrivacy`, and other regulatory-focused subreddits such as `r/LegalAdviceEU`. Direct outreach was limited to users engaged in GDPR-related discussions. These communities bring together diverse stakeholders, including developers, legal and privacy experts, and end users, whose interactions shape how compliance is interpreted and applied in practice.

Prior to participation, users were required to agree to an IRB-approved consent form that describes eligibility, data privacy, and compensation terms. After one week, we stopped data collection and deleted the flyer from Reddit. As disclosed in the consent form, the first 100 participants, regardless of their responses, received a $5 Amazon gift card upon completion of the ~8–12 minutes survey (equivalent to ~$25/h). To maintain anonymity, those who wished to be compensated entered their email addresses in a separate, unlinked form at the end of the survey. Not all completed this step.

3.1.2 Survey Analysis Process. A power analysis [11] indicated that approximately 100–120 total responses would provide at least 80% power to detect medium effects (e.g., $\chi^2 = 0.30$, $\rho = 0.25$, $f = 0.25$). Based on this, we estimated that around 150 submissions would yield more than 80% power. To reach this target, recruitment (described in Section 3.1.1) was conducted over one week. We removed the public survey post and stopped direct outreach after this period.

We collected 522 submissions. Given the open recruitment strategy on Reddit and the compensation offered to the first 100 participants, we anticipated a substantial number of low-quality or automated responses. We therefore implemented a conservative multi-step validation process prioritizing data integrity over sample size. Instead of attention-check questions, we employed Qualtrics'

built-in tools to flag low engagement, e.g., straight-lining, bot-like speed, and duplicate entries. We also manually reviewed all 522 responses regardless of flagging for poor-quality indicators, such as short completion time (submitting < 1 minute), repeated text, nonsensical or blank open-ended answers, skipped or incomplete question responses, and characteristics consistent with AI-generated text [40, 63]. Submissions missing only the optional demographic questions were retained. This process resulted in 223 valid responses, of which 80 included completed demographic information.

In the *quantitative analysis*, we used the Chi-squared test [23] to determine whether there is a significant correlation between two categorical variables (e.g., whether reported GDPR challenges differed by company type, size, or location, and job role). For the Likert-scaled questions, we used the Kruskal-Wallis test [6]. When both variables were ordinal, we applied Spearman's rank correlation [54] to evaluate monotonic relationships. To analyze repeated-measures data, such as how responses to multiple Likert items vary by participant background, we used a Mixed ANOVA [26], treating each item as a within-subject factor and demographic variables (e.g., job role and education) as between-subject factors. We also applied a LASSO logistic regression model [64] to identify which credibility strategies best predict users' decision-making behavior.

We analyzed the *open-ended survey questions* (e.g., privacy certifications (Q6) or steps to assess credibility (Q24)) using an open-coding procedure to identify categories based on prevalent themes in the data, until saturation was reached. We initially assigned codes to each response based on keywords, common terms, or themes. For example, if the response to Q24 was "check profile" or "cross-fact check", we assigned initial codes: check user profile or seek external verification. We continued until no new codes were introduced. We then grouped similar codes to form broader categories and refined them iteratively during annotation to ensure clarity, internal coherence, and to reduce overlap. We then examined the "Other" code to see if a new theme emerged. After finalizing the categories, a privacy expert (i.e., the last author) reviewed the codes, categories, and responses to verify their accuracy and consistency.

3.2 Reddit Post Analysis Approach

We cross-validate the survey findings by analyzing posts from `r/gdpr` and `r/europrivacy`.

3.2.1 Data Collection. We collected posts related to privacy and software development from the `r/gdpr` and `r/europrivacy` subreddits using the Python Reddit API Wrapper [5] in April 2025. We did not filter posts based on upvotes or comment count. Although these subreddits focus on privacy regulations, they often include content unrelated to privacy in software development, like career advice, news, surveys, or published articles. Therefore, we built a search term list similar to the Hark method [25], which draws on two commonly used and complementary privacy taxonomies: Solove's taxonomy of privacy harms [52] and the privacy-enhancing technology taxonomy [66]. Each privacy category serves as a base search term. We expanded these base terms with relevant synonyms to broaden coverage, resulting in a final set of 38 search terms (see Appendix K.1 - Table 15 for more detail). We then queried Reddit for these terms and retrieved a total of 2,248 posts (i.e., 2,014 from `r/gdpr` and 234 from `r/europrivacy`), where the majority of posts were created after January 1, 2020.

3.2.2 Ground-Truth Dataset Creation. Our goal is to identify privacy challenges in software development by analyzing the extracted 2,248 posts. To achieve this, we first manually annotate a subset of posts related to both *privacy* and *software development* with the type of GDPR-related challenges to establish a ground truth. We then use a large language model (LLM) to automate the annotation process. Specifically, we randomly select a subset of posts from the collected Reddit data and label them in two steps: (1) determining whether each post is relevant to both privacy and software development; and (2) categorizing the relevant posts according to the GDPR-related challenges or questions they addressed.

In Step 1, two authors with expertise in privacy and software development independently reviewed a subset of posts across multiple rounds to determine whether they were related to both *privacy* and *software development*. We initially randomly selected 100 posts and repeated the process until we collected at least 100 posts relevant to privacy and software development. In each round, we calculated the inter-rater reliability agreement statistic, Cohen's Kappa. This value ranged from 0.684 to 0.834, depending on the round, indicating "substantial" to "almost perfect" agreement. After each round, a third expert reviewed the posts with disagreements and documented their own response. All three annotators then met to resolve disagreements before proceeding to the next round. After four rounds, we identified 116 posts out of 350 posts related to both privacy and software development (See Appendix K.2 - Table 16).

In Step 2, the same two annotators independently reviewed these 116 posts and assigned one or more labels with a set of GDPR challenge categories, accompanied by their rationale. We followed a hybrid grounded analysis [27], where the initial categories were adopted from CIPL's 2017 report on GDPR implementation [8]. However, we expanded the categories as new themes emerged. If the post did not match the initial challenge, the annotators first labeled them as "Others". They then analyzed the "Others" category to identify recurring themes, which led to three additional GDPR challenge labels, namely, "consent", "GDPR definitions", and "data processing agreement (DPA)", in Round 1. In this round, they also agreed on 79/116 posts. Thus, the annotators relabeled the entire dataset again using the updated list of challenges, achieving agreement on 98/116 posts. A third expert reviewed the disagreements and made a final decision. Appendix L - Table 20 presents the initial challenges, along with those added during our analysis.

3.2.3 Automated Post Labeling. We automate labeling of the remaining 1,898 posts using in-context learning with LLMs [7] and a human-in-the-loop (HiTL) approach [44]. In HiTL, the ground truth dataset is used to train and evaluate a model, which is then used to label a larger, unreviewed dataset. The two dataset distributions are assumed to be very similar because the ground truth data was randomly sampled from the same dataset as the larger set. If recall on the testing evaluation is high (i.e., $tp/(tp + fn) \geq 0.85$), then HiTL can reduce the investigator's effort to reject false positives without significant losses due to false negatives. For example, if 30% of 1,898 posts are privacy and development related, then we may miss ~85 out of 569 posts as false negatives, while reliably adding

Table 2: F1 Scores for Different Models Using CoT

Model	Privacy	Development	Challenges
Llama3.2-1B-Instruct	0.8560	0.4030	0.3584
Llama3.2-3B-Instruct	0.8270	0.4211	0.2939
Llama3.1-8B-Instruct	0.9189	0.5405	0.4768
Qwen2.5-7B-Instruct	0.8710	0.4407	0.6388
gpt-3.5-turbo-1106	0.9498	0.6477	0.4019
gpt-4o-mini	**0.9535**	0.4552	0.6683
GPT-5-mini	0.9498	0.8008	**0.7357**
GPT-5	**0.9423**	**0.8531**	0.6857

484 additional posts through manual validation. Although unsupervised machine learning (ML) approaches, e.g., topic modeling, could be used to cluster posts and challenges, it is best applied when we do not have a priori knowledge. In this work, we instead begin with developer-reported privacy challenges from a survey and seek to identify evidence of these challenges in the Reddit posts, which requires supervised ML to classify texts into challenge categories. To that end, we supervise prompt design with demonstrations for use with in-context learning to classify the posts.

We designed two text classification tasks: (1) separately classify posts that are related to both privacy and development, and (2) classify posts by the privacy compliance challenges raised. For both tasks, we first refined the prompts on a subset of the ground-truth dataset (posts for the development set) before testing them on the remaining subset of unseen posts (the holdout set). We also retain a subset of data for any potential prompt refinements.

For the first classification task, we selected 100 posts (~30%) from the ground truth as the development set and evaluated performance across both open-source and proprietary LLMs. Specifically, we tested three LLaMa, one Qwen, and four OpenAI models using three prompt engineering strategies: zero-shot prompting, 2-shot learning [7], and chain-of-thought (CoT) prompting [67] – in total 24 experiments. Our final prompt (i.e., Appendix K.3 - Listing 2) performed consistently the best across both labels and all models.

We then randomly sampled an additional 140 posts (~40%) not used in the development set to test the final prompt across all models to select the best model. While GPT-4o-mini achieved the highest $F1 = 0.9535$ for the privacy label, GPT-5 produced the best overall results, with the highest $F1 = 0.8531$ for the development label and a comparable $F1 = 0.9423$ for the privacy label, which is 1.2% below the highest score in privacy (see Table 2). Thus, we adopted GPT-5 with CoT for the HiTL labeling effort.

After running our prompt on the entire dataset using GPT-5, 591 additional posts were identified as related to both privacy and software development. We manually reviewed and updated the predicted labels and removed 41 irrelevant posts. Overall, out of the 2,248 extracted posts, 666 posts (30%) were identified as related to privacy and software development (550 labeled via HiTL and 116 from the ground truth). We did not review the posts labeled as unrelated. However, since recall was high (i.e., 0.9286 for privacy and 0.8571 for development), we likely captured most related posts.

We then applied the same process to identify the *expressed challenges* (i.e., the second classification task). We first extended the set of 116 posts related to privacy and development by manually annotating an additional 38 posts randomly sampled from the HiTL method. This yielded 154 posts labeled with challenges for use in development and holdout sets, and to test our final prompt. The details for testing prompts and models are in Appendix K.3. After running the experiments with the three prompting strategies on the same eight models, we found that the CoT prompt (see Listing 1) performs best. We then used our final prompt to test the models on the 62 unseen labeled posts. As shown in Table 2, GPT-5-mini achieved the best performance with an $F1 = 0.7357$. Although the F1 score is not high, it is comparable to prior work for identifying privacy concepts in policy text [4]. Furthermore, research [36] shows that manually classifying posts is more labor-intensive and error-prone than verifying and correcting labels. Thus, we adopted GPT-5-mini with CoT for identifying GDPR challenges in the dataset.

We then ran the final prompt on the remaining 512 posts and observed that the model assigned exactly one label per post (i.e., no post lacked a label). Next, the two annotators reviewed all posts with their assigned labels and flagged those that were incorrect. They reviewed the incorrect posts with the third privacy expert, which led to changes in the labels for 138 posts. Finally, the same two annotators reviewed all posts labeled as "Others" for potentially new themes. They identified two additional challenge categories that were not present when constructing the ground truth: "data transfer" and "generic compliance" (see Appendix L - Table 20 for their definition). A senior privacy expert reviewed these final mappings to ensure their accuracy.

Listing 1: CoT Prompt for Detecting GDPR Challenges

```
You are an expert system that needs to determine if Reddit posts
      are in any way related to various GDPR challenges and to
      provide an explanation for your reasoning.

These are the only valid challenges and their definitions:
{labels' list}

Step 1: Review the examples of posts and their answers.
{examples' list}

Step 2: Understand how to apply this reasoning to new posts, using
       the examples and explanations as guidance.

Step 3: Apply this understanding to the following post.

Step 4: Remember to prioritize accuracy and clarity in your
      analysis, using the provided context and your expertise to
      guide your evaluation. If you are uncertain about the
      classification, choose 'Others' and explain your rationale
      for this uncertainty.

Step 5: Determine the correct classification for the following
      post. Make sure to review your response and provide a
      rationale for your selection:

For the following post:
- Return a valid array of containing ONLY the most applicable
      label from the list above that applies to the post.
- Do NOT ever include more than one label.
- Do NOT include the definition of the label in your answer, only
      the label name in quotes.
- If none apply, return an empty array, [].
- Do NOT include any code, only the array and your explanation.

Now label the following post.
Title: {post title}
Text: {selftext}
Explanation:
```

4 Ethical Considerations and Limitations

Ethical Considerations: The survey and post analysis were conducted under our Institutional Review Board's review and approval. All participants were informed about the research, the risks and benefits of participation, compensation, and confidentiality before providing their consent. Participation was voluntary and withdrawable at any time. We did not collect personally identifiable information, and we have measures in place to ensure the anonymity, confidentiality, and security of responses. While contact information for all investigators and the IRB team was provided, no participants contacted them about the study or the compensation. Prior to analyzing the publicly available Reddit posts, we removed all usernames and comments and only shared post titles and body text with LLMs.

Limitations: We observed limitations in our research design that we attempted to mitigate. While we used screening questions to target individuals in privacy or software development, our sample may not fully represent all privacy practitioners or developers engaging with GDPR. Reddit users participate voluntarily, which may limit the representativeness of our sample. To reduce self-selection bias, we deliberately avoided advertising the survey as "privacy research". However, the survey remains subject to self-report bias, recall bias, and social desirability bias — e.g., by overreporting certifications or involvement in GDPR compliance. Despite efforts to validate submissions, including the use of Qualtrics' bot detection and manual screening, some responses may have been generated or influenced by AI tools. We conservatively excluded suspicious entries, which may have inadvertently removed legitimate responses.

We used a 4-point Likert scale to reduce response bias, including central tendency (avoiding clear choices) and social desirability bias (giving socially acceptable answers), as respondents often select the midpoint when they are unsure or wish to avoid disagreement [10, 21]. Our survey used individual agreement-based items rather than a composite Likert scale that aggregates multiple statements measuring a single construct. This approach allowed us to capture perceptions on specific topics while minimizing bias.

We allowed skipping questions and avoided emotionally charged wording to minimize recall bias. Some items, e.g., *Q10: "Within my company, the knowledge about privacy compliance is up to date"*, may be perceived as leading but were intended only to assess organizational awareness rather than individual knowledge. To strengthen our findings' validity, we conducted statistical analyses to identify significant relationships. Nonetheless, these limitations may affect results' accuracy; we thus advise caution in interpreting them.

We used predefined challenge categories from the 2017 CIPL GDPR implementation report [8] in our survey, as well as an open-ended option to capture newly emerging categories. However, requiring participants to type new challenges may have led to inconsistent capture of new categories, unlike in the post analysis.

While our findings highlight patterns in two regulatory-focused subreddits, they may not be generalizable to other communities or regulatory contexts, such as the CCPA, which future work could address. Although our keyword-based search approach used terms from comprehensive privacy taxonomies, some relevant posts may have been missed. Since these terms and their synonyms cover the major privacy themes, most relevant posts were likely captured.

5 Survey Findings

In this section, we report our findings related to RQ1-RQ3.

5.1 Demographics, Professional Background, and Reddit Interaction Patterns

Although we had 223 valid participants, we received only 80 responses to the optional demographic questions (a 64% non-response rate) as they were placed at the end of the survey. Appendix C - Table 13 presents the demographic findings: the majority of respondents identified as male (73%, n = 58) and are mostly between the ages of 35-44 and 25-34, representing 46% (n = 37) and 40% (n = 32) of respondents, respectively. Professional degrees (i.e., certificate degrees) are held by 39% (n = 31) of participants, followed by 34% (n = 27) who have completed graduate studies. The majority hold degrees in computer science (35%, n = 28), data science (23%, n = 18), and information technology (21%, n = 17). 30% (n = 24) work in companies with 101+ employees, while 51% (n = 41) work in medium-sized companies (21-100 employees).

In contrast, for professional background questions (i.e, *Q1-Q8*), we received 223 responses, as shown in Table 3. Most participants work in for-profit companies, half are based in North America (although the subreddits primarily discuss EU laws), and have 3–5 years of job experience. About 80% report involvement in GDPR-related tasks or compliance. Among the roles, Technical Roles have the highest representation (48%), followed by Management and Executive Roles (23%), Others (16%), Academic and Research (9%), and Legal and Compliance (4%) (see Appendix D - Figure 5a). We further break technical roles into more granular categories. As shown in Appendix D - Figure 5b, within the technical roles group, 40% are in "software development", 32% in "data science", and 14% each in "cybersecurity" and "IT operations" roles.

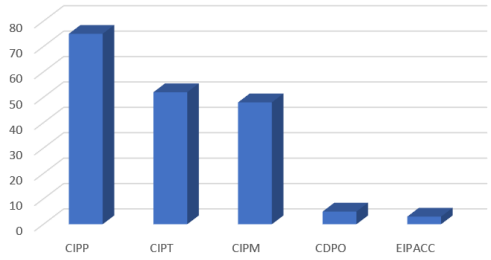
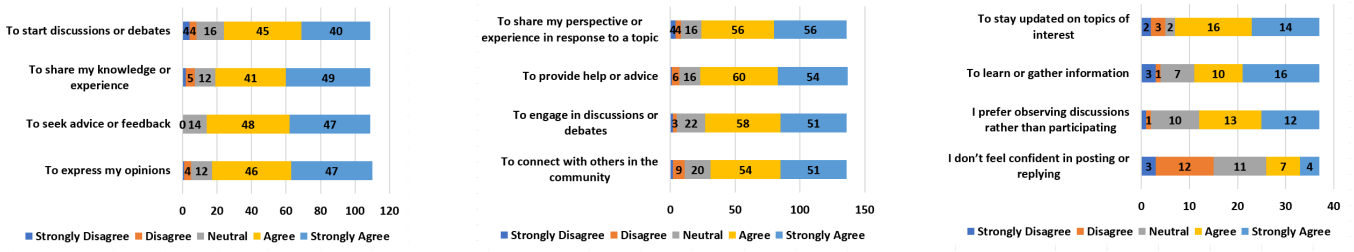


Figure 2: Top Five Privacy Certifications Categories

We evaluated whether participants held any professional certifications. We categorized the 195 valid responses into six groups. Privacy-related certifications dominate the responses, with ~90%, whereas the other categories were mentioned only a few times (See Appendix E - Figure 6). We further analyzed the privacy certifications in more detail. Participants listed certifications such as CIPP (41%), CIPT (28%), CIPM (26%), CDPO (3%), and EIPACC (2%) (see Figure 2). Most certifications are issued by the International Association of Privacy Professionals (IAPP) [32], which is particularly popular in North America. EU-specific certifications such as CDPO [46] and EIPACC [16] were less frequently mentioned, likely reflecting the geographic distribution of respondents.

Table 3: Professional Background of the Participants

Category	Details
Company Type	For-profit (51%), Non-profit (10%), Educational (15%), Government (8%), Research (9%), Self-employed (6%), Other (1%)
Company Location	North America (50%), Europe (14%), Central America (13%), South America (13%), Africa (6%), Other (4%)
Years at Current Job	0-2 Years (7%), 3-5 Years (45%), 6-10 Years (41%), 11+ Years (8%)
Years in Privacy/Data Roles	No experience (4%), 1-2 Years (21%), 3-5 Years (50%), 6-10 Years (23%), 11+ Years (3%)
GDPR Involvement	Yes (80%), No (14%), Unsure (4%), Prefer not to say (2%)



(a) Motivations for Create Posts (Q19) **(b) Motivations for Post Replies (Q21)** **(c) Motivations for Only Read (Q23)**

Figure 3: Motivations For Interaction on Reddit

We explore whether participants interact with Reddit and, if so, what the nature of their interaction is. As shown in Appendix F - Figure 7, out of 223 respondents, ~75% reported that they engage with GDPR-related posts at least a few times a week or daily, similar to the findings in [48]. ~8% of users almost always post or reply about GDPR topics, while about 50% have posted or replied about GDPR, frequently in the last six months (see Appendix G - Figure 8). The most common behavior among participants is to equally post and reply to others (34%), followed by mostly replying to others' posts (31%) and primarily creating posts (19%). About 17% mostly read posts without posting or replying (see Appendix H - Table 14).

5.2 Users' Motivation for Interaction

To address *RQ1*, we explore the motivations for engaging in discussions on the subreddits and examine whether these motivations relate to users' professional, organizational, and privacy-related backgrounds. Depending on the interaction types (see Table 14), we asked tailored follow-up questions *(Q18–Q23)*.

Users' Motivations for Creating Original Posts: 78 participants described their motivation for creating original posts on GDPR-related subreddits (including those who equally create or reply). We noticed that most mentioned they post to share their knowledge or experience (37 responses, 47%), express opinions (19 responses, 24%), seek advice or feedback (10 responses, 13%), and start discussions or debates (8 responses, 10%). For example, one respondent said *"I create posts on the importance of security"*. Several noted that *"they post when they consult external sources for privacy or legal compliance and want to share those findings with the community"*. About 5% are labeled as "Other", as they do not fit into any categories.

We then asked them to rate their motivations on a Likert scale to validate the open-ended responses. We received 110 responses. 83% agreed or strongly agreed that they are motivated by a desire to share their knowledge or experience (see Figure 3a). Seeking advice or feedback and expressing their opinions are also key motivators for 86% and 85% of respondents (at least agreeing), respectively. Finally, 78% agreed or strongly agreed that starting discussions or debates encourages them to share original content. These findings are aligned with the open-ended responses.

Users' Motivations for Post Replies: 87 participants expressed their motivation for replying to others' posts. Most replied to provide help or advice (39 responses, 45%), share their perspective or experience (27 responses, 31%), engage in discussions or debates (9 responses, 10%), or connect with others in the community (5%). For example, one reported *"providing help and advice, sometimes engaging in discussion"*, while some said they specifically aim to challenge or validate claims, e.g., *"cross-check with multiple sources before replying"*. 9% were labeled as "Other", (e.g., motivations like *"telling others how impressed I am"*) which did not fit our themes.

138 participants ranked their motivation on a Likert scale. The majority (83%) agreed that they are motivated to provide help or advice. They also (strongly) agreed with: sharing their perspectives or experiences (82%); connecting with others in the community (77%); and engaging in debate (80%) (see Figure 3b).

Users' Motivations for Reading without Engaging: We received 19 responses from those who mostly read the posts without posting or replying. Nearly half (9 responses, 47%) reported reading primarily to learn or gather information. Other motivations include staying updated on topics of interest (3 responses, 16%), preferring to observe discussions rather than actively participate (3 responses, 16%), and lacking confidence in posting or replying (2 responses, 11%). Two responses were labeled as "Other." Examples include *"I read and analyze to stay up to date with emerging trends"* and *"learn about it first and then share it with people around you"*.

A total of 37 participants responded to the Likert-scale prompt. As shown in Figure 3c, the most common motivations are staying updated on topics of interest (81%), learning or gathering information (70%), and preferring to observe discussions rather than

participate (68%). A smaller proportion (30%) indicated they read posts because they do not feel confident in posting or replying.

We also examined whether the types of motivations relate to users' job roles (H1a), privacy-related backgrounds, i.e., privacy-related certifications (H1b), and privacy-related experience (H1c), and company size (H1d). We did not observe any statistically significant relationships between job roles (N = 27–118, depending on the motivation) and their motivations. To evaluate (H1b), we grouped participants into two categories: those with privacy-related certifications and those without. Participants holding *privacy-related certifications* reported significantly stronger motivations for creating posts to share knowledge and experiences ($U = 956.50$, $p - value = 0.001$, $r = 0.476$, $N = 94$) and to express opinions ($U = 859.50$, $p - value = 0.032$, $r = 0.310$, $N = 95$) compared to others. We also observed significant relationships between *privacy-related experience* and motivations for creating posts to share knowledge and experiences ($H = 14.30$, $p - value = 0.006$, $\epsilon^2 = 0.112$, $N = 95$) and replying to share perspectives or experiences ($H = 8.80$, $p - value = 0.032$, $\epsilon^2 = 0.051$, $N = 117$). Company size was associated with motivations for seeking advice and feedback ($H = 10.96$, $p-value = 0.027$, $\epsilon^2 = 0.279$, $N = 30$) when creating posts and replying to share perspectives or experiences ($H = 9.02$, $p-value = 0.029$, $\epsilon^2 = 0.140$, $N = 47$). These findings suggest that privacy expertise and organizational context may influence motivations for participating in GDPR-related online discussions.

Summary: Regulatory-focused subreddits support both advice-seeking and knowledge-sharing around compliance. Privacy-related experience, certifications, and organizational context may shape how users participate in these discussions.

5.3 Perceptions and Challenges

We examine GDPR challenges and perceptions of compliance across professional and organizational backgrounds (i.e., *RQ2*).

We asked participants to describe the aspect of GDPR compliance they find most challenging, with response options adopted from CIPL's 2017 report on GDPR implementation [8]. We received 220 responses. Multiple options were allowed (see Table 5.3). 121 respondents (55%) considered data protection impact assessments (DPIAs) as the most challenging. Data breach notification is the second-most reported challenge (49%, 107 responses), followed by data subject access requests (DSARs) (39%, 85 responses), and maintaining records of processing activities (RoPA) (29%, 64 responses). Only 1% (3 responses) selected "Other, please specify" which included *"data retention issues"*, *"designing compliant personal data flows across cloud systems and organizations"*, and *"the whole process and divergence of opinions making arbitrations difficult"*. The results suggest that the processes for minimizing risk are perceived as the most challenging aspect of GDPR compliance.

We also explore whether participants' company characteristics, i.e., type (H2a), size (H2b), and location (H2c), job roles (H2d), as well as holding privacy-related certifications (H2e), influence their perception of these challenges (See Appendix B.2). While we find significant relationships between company type or size, and holding privacy-related certifications and GDPR challenge selection ($N = 190$) with ($p - value = 0.0001$, $Cramr'sV = 0.2123$), ($p - value = 0.0195$, $Cramr'sV = 0.1564$), and ($p - value < 0.001$, $Cramr'sV = 0.2819$), the company's location or job roles show no such relationship (see Appendix B.2 - Table 9). These results indicate that although companies of different types and sizes focus on various compliance issues and face distinct regulatory concerns, within the company, the focus on compliance issues is homogeneous across management, legal, and development teams, except for those holding privacy-related certifications. For example, for-profit companies often selected challenges such as data breach notification and DPIAs, while government and academic groups emphasized these relatively less and highlighted other concerns. Whether the company is located in North America or the EU does not have an impact, likely because the focus is only on GDPR challenges.

Table 4: The Distribution of GDPR Challenges

Category	% (Count)
Data Protection Impact Assessment (DPIA)	55% (121)
Data Breach Notification	49% (107)
Data Subject Access Requests (DSARs)	39% (85)
Records of Processing Activities (RoPA)	29% (64)
Other	1% (3)

Next, we check whether the participants' organization has any in-house legal team specializing in privacy compliance. The majority (62%, 137 respondents) indicated that their organization has a dedicated privacy or legal team. An additional 25% (56 respondents) reported having a legal team that is not explicitly dedicated to privacy. However, 10% noted the absence of an in-house legal team, while 2% were unsure. These findings show that while 87% of organizations have some form of in-house legal support, only 62% have teams dedicated explicitly to privacy compliance. This limited in-house dedicated privacy teams suggests that many organizations may need supplementary support. This is further reflected in participants' reports of whether they or their companies seek external privacy or legal advice. As shown in Appendix I - Figure 9, only 13% (29 respondents) said they manage all compliance internally, while 9% (21 respondents) rarely seek outside help, and 5% (10 respondents) were unsure. The majority (73% of participants) report seeking external privacy or legal advice at least occasionally. We also notice that those in small companies (i.e., < 100) reported higher reliance on external legal or privacy advice (i.e., 80%). These results indicate that companies routinely compensate for gaps in internal privacy expertise by engaging external specialists or other tools, highlighting an underlying mismatch between regulatory demands and available in-house capabilities. The results are also aligned with findings in [48] where a privacy officer role was not often present, even in larger companies, and ~50% seek external advice, check the best practices like NIST guidelines, or use forums.

We assessed whether participants perceive their company's privacy compliance knowledge to be up to date. A total of 220 responded. Most expressed confidence in their organization's privacy compliance knowledge, with 86% selecting "agree" or "strongly agree". 10%, were neutral, while only 3% (strongly) disagreed.

We then evaluated whether their confidence in their company's privacy compliance is influenced by the presence of internal legal

Table 5: Primary GDPR Information Sources

Source of Information	% (Count)
Official regulatory websites (e.g., ICO)	57% (126)
Professional networks and forums	51% (112)
Industry conferences and webinars	43% (95)
Legal and compliance publications	36% (79)
Internal corporate training	24% (53)
Other	1% (3)

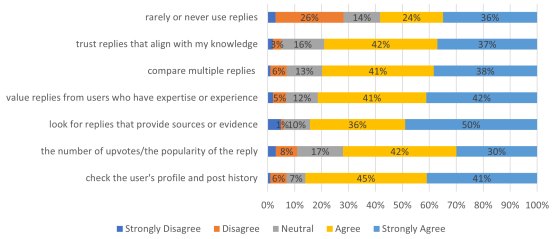


Figure 4: Credibility Evalution Strategies

resources (H3a) and reliance on external legal advice (H3b) (See Appendix B.3 - Table 10). With $H = 8.4773$, $p - value = 0.0371$, $\eta^2 = 0.0294$, $N = 190$, we find a significant relationship for H3a, but did not observe a statistically significant relationship for H3b, with $\rho = 0.0083$, $p - value = 0.9096$, $N = 191$, highlighting that a dedicated privacy or legal team increases the likelihood that employees perceive their company's privacy knowledge as up to date (similar to [48]), but this confidence is largely consistent, regardless of whether and how often they relied on outside legal support.

We received 221 responses regarding their primary sources of information for staying updated on GDPR. As shown in Table 5, the most common sources are official regulatory websites (57%, 126 respondents) and professional networks and forums (51%, 112 respondents), suggesting that they prioritize authoritative resources and peer communities. Industry conferences and webinars (43%, 95 respondents) and legal and compliance publications (36%, 79 respondents) are also considered. Only 24% cited internal corporate training, indicating its limited role in ongoing education, aligned with findings in [30]. 3 respondents selected "Other." We examined whether primary sources of GDPR knowledge vary based on their job roles (H4a), company size (H4b), educational background (H4c), and privacy-related certifications (H4d) (See Appendix B.4 - Table 11). We do not find statistically significant relationships between GDPR knowledge sources and job roles, company size, or educational background. However, we observed a significant relationship for privacy-related certifications (H4d), with $p - value = 0.0046$, $Cramr'sV = 0.2820$, $N = 191$. These findings suggest that the sources of GDPR information do not differ substantially across professional, company, or educational backgrounds. However, users with privacy-related certifications may rely on different sources.

Summary: Participants consider DPIAs and data breach notifications the most challenging aspects of GDPR compliance. Challenges vary by company type and size, and holding a privacy-related certification. Most companies rely on external legal help to fill gaps in their internal privacy expertise.

5.4 Credibility Evaluation & Ambiguity Resolutions Strategies

We answer *RQ3* by exploring how participants evaluate the credibility of replies or how they disambiguate before responding.

Participants who created posts were asked how they decide whether the reply is credible. We manually categorized the open-ended responses into eight groups, as shown in Table 6. For example, 22% mention they check *the user profile and history*, e.g., "verify profile and previous posts including replies." 18% of them seek *external verification* including "Google and fact checking website" and "reputable news articles across vast news outlets", while 17% look for *sources/evidence* through strategies such as "check original post, look for the source citation" or "look for supporting evidence in the response." Smaller portions (8%, 7% and 6%, respectively) search for *community feedback*, e.g., "review comments" and "observe feedback from other users," just use the *responses directly* e.g., "I use them to make decisions directly related to the topic," or focus on *critical thinking or logic*, like "I take a minute to understand what the topic is all about." Some participants use *multiple tactics*, such as "evaluate subreddit and poster's history, check for verified badges, look for sources, search for additional evidence, and crosscheck externally." ~11% do *minimal verification* or *were unsure*. Figure 4 shows the participants' rating of their agreement with seven strategies for credibility assessment. Most agreed with relying on replies that include evidence (86%) or checking user profiles and history (86%) while other strategies were less commonly agreed on. Overall, users prioritized expertise and users' history over popularity, but a small number consider minimal verification.

We used mixed ANOVA analysis to examine whether participants from different professional backgrounds, including job role (H5a), educational background (H5b), company size (H5c), and privacy-related certifications (H5d), differed in how they evaluated the credibility of replies and whether they preferred different credibility evaluation strategies. Detailed statistical results are provided in Appendix B.5 - Table 12. The between-subject groups are participants' job roles, company sizes, educational backgrounds, and privacy-related certification status, while the within-subject factor is the set of credibility evaluation strategies. A main effect of strategy indicates that participants generally preferred some credibility evaluation strategies over others. An interaction effect indicates that different groups preferred different credibility evaluation strategies. For job role ($N = 94$), we found a significant main effect ($p - value = 0.0498$), but no interaction ($p - value = 0.127$), suggesting that while roles differ in overall agreement, they rank the strategies similarly. For education ($N = 29$), neither the main effect ($p - value = 0.719$) nor the interaction ($p - value = 0.925$) was significant, indicating little influence on strategy use. For company size ($N = 29$), the main effect was not significant ($p - value = 0.126$), but we observed a significant interaction ($p - value < .001$), indicating that strategy preferences varied by company size. For privacy-related certifications ($N = 94$), we did not observe a significant main effect ($p - value = 0.452$) or interaction between certification

Table 6: Strategies for Verifying Responses

Category	% (Count)
Check user profile/history	22% (15)
Seek external verification	18% (13)
Look for sources/evidence	17% (12)
Community feedback	8% (6)
Use response directly	7% (5)
Critical thinking or logic reasoning	6% (4)
Comprehensive strategy (multi-step)	11% (8)
Minimal / Don't know / Unclear	11% (8)
Total	71

status and strategy ($p - value = 0.569$), indicating that participants with and without privacy-related certifications ranked credibility evaluation strategies similarly. Overall, while privacy, professional, and organizational factors may influence credibility evaluations, most users rely on similar strategies regardless of background.

Participants who mainly replied to others were asked how they respond when encountering an unclear or incomplete post. Their open-ended responses showed various strategies. ~50% mentioned *seeking clarification*. E.g., one participant reports that *"politely ask clarifying questions, provide construction feedback, and share relevant insights"* while another said *"request for a complete post from the users post"*. ~13% reported they *avoid replying to unclear posts* where one said *"I don't respond to unclear post that I haven't understood"*. A small group used other strategies like *"sending a private message to the subreddit"* or *"say so and elaborate on my interpretation of the question."* Overall, users relied on clarification, interpretation, and flexible responses when dealing with ambiguous posts.

We used LASSO logistic regression [64] to determine whether credibility assessment strategies can predict how users decide on the responses. For "those who use the responses directly to make decisions", the model selected three strategies: checking upvotes, checking the user's profile and history, and trusting replies that match the user's existing knowledge, with coefficient values of 0.422, 0.385, and 0.049. These users appear less focused on validating sources and tend to be less critical. The model did not select any strategy as a predictor for "those who use the responses as a reference, but verify the information independently", suggesting that independent verification is not explained by credibility strategies in our survey. The model selected all seven strategies for "the users who use the response as part of many opinions to form their view", but none of the predictors showed a significant effect, which indicates they rely on many small strategies together, rather than one or two major ones. For "those who mainly ignore responses unless they are highly voted or come from a credible source", two predictors were significant (i.e., pay attention to upvotes or rarely use responses with coefficient values of 0.791 and 0.641). Lastly, for "those who only want to begin a discussion or ask follow-up questions", the model selected four predictors with coefficient values of 0.700, 0.537, 0.498, and 0.025, where three are statistically significant. These findings suggest that users rely on different credibility strategies depending on how they use Reddit replies.

Table 7: GDPR Challenges in Annotated Reddit Posts

* Indicates GDPR challenges also identified in the survey.

Category	% (Count)
Cookie Consent	20% (132)
General Consent	18% (119)
Data Processing Agreements (DPAs)	12% (81)
Data Transfer	12% (77)
Other	9% (62)
GDPR Definitions	8% (55)
*Data Protection Impact Assessment (DPIA)**	6% (39)
*Data Subject Access Requests (DSARs)**	6% (38)
Generic Compliance	5% (32)
*Records of Processing Activities (RoPA)**	4% (24)
*Data Breach Notification**	1% (10)
Total GDPR Challenges	100% (669)

Summary: Most users assess credibility using evidence-based methods like verifying external sources, checking profile histories, or upvotes, and seek further clarification for ambiguous posts; yet, a smaller number never conduct verification or look for sources. Selecting one strategy over another is explained by some professional backgrounds, but privacy experience does not affect such decisions.

6 Reddit Posts Analysis Results

In this section, we answer *RQ4*. The Reddit post analysis described in Section 3 mapped each post to one or more GDPR challenge categories. Table 7 shows the distribution. Among the 666 posts, consent was the challenge most frequently discussed (38%), followed by data processing agreements (DPAs) (12%) and data transfer to a third country (12%). References to GDPR definitions, DPIAs, and DSARs also appeared regularly. Less frequent challenges included generic compliance, RoPA, and data breach notifications.

6.1 Reddit Post Analysis vs. Survey Findings

6.1.1 GDPR Challenge Overview. Our Reddit post analysis complements and extends the survey findings, and also reveals notable differences in concerns. While DPIAs (55%) and data breach notification (49%) are the most frequently cited challenges in the *survey* (see Table 4), Reddit discussions focused more on consent (38%), particularly cookie consent, which accounted for ~53% of all consent-related posts. This difference is mainly due to the broader scope and open-ended nature of Reddit discussions, reflected in the high proportion of posts labeled as "Others" (9%) and six additional categories absent in the survey. The challenges that overlap are shown in *italics* with *. As shown in Table 7, the first overlapping challenge, i.e., DPIA, is the seventh most common challenge in the post analysis. Our analysis reveals that only 111 of the 666 posts (17%) matched challenge categories explicitly included in the survey. This result confirms that the survey captured only a subset of the real-world GDPR challenge space and that practitioners encounter a wider range of issues in practice than those reported by CIPL [8]. When we filter posts to include only those labeled with survey categories (Appendix J - Figure 10), the distribution of challenges

mostly matches, with DPIAs, DSARs, and RoPA appearing in the same relative order across both datasets. The only exception is data breach notifications, which are rarely mentioned in the posts, likely due to the risks of publicly disclosing such incidents [18, 19, 53].

6.1.2 Interaction Types. We analyzed whether posts in each category contained follow-up responses from the poster. Our analysis showed that 83.33% of RoPA posts had follow-up comments from the original poster, followed by 78.18% of GDPR definitions, 75% of generic compliance, and 70% of breach notification posts. These comments mostly covered questions and clarifications, suggesting the complexity of RoPA challenges.

6.1.3 Analysis of Regulatory Sources. We also evaluated which official documents or GDPR articles were most cited in responses. ICO and EDPB were cited the most (i.e., 165 and 118 times), while IAPP and NIST sources were rare (unlike the survey findings, where NIST was more commonly used). As for GDPR, Art. 28 (Processor) has been cited the most, followed by Art. 6 (Lawfulness of processing), Art. 13 (Information to be provided where personal data are collected from the data subject), and Art. 3 (Territorial scope). Articles less frequently discussed include Art. 30 (RoPA), Art. 44 (General principle of transfers), and Art 46. (Transfers subject to appropriate safeguards); even though 16% of challenges were tied to them. Art. 35 (DPIA) was only directly referenced once, potentially due to users using the term "DPIA" rather than the article number. We identified 582 (11.3% of all comments) total references to specific GDPR articles within the comments in our dataset.

6.2 Analysis of GDPR Challenges

6.2.1 Cookie Consent. Most posts in this category are concerned about what can be tracked without asking for consent, the difference between strictly necessary cookies and other types, and confusion about GDPR vs. the ePrivacy Directive (ePD). A poster asked *"if consent was required for cookies implementing first-party analytics."*

A large portion of the posts discuss Google Analytics (GA) and the consent requirements for using it. One poster sought *"May I store GA cookies without consent under the (ePD)?"* Some of the posts also asked whether a personal blog requires cookie banners. Documenting and storing consent for audit purposes was another recurring question. The responses mainly focused on whether tracking is strictly necessary or not. However, there was occasional confusion regarding "strictly necessary" vs. "functional" cookies. While some users claim that consent is not required for functional cookies, one responder mentions *"The author of the page you're citing has invented the distinction between strictly necessary and functional cookies out of thin air. [..] this distinction has no basis in the actual laws or in authoritative guidance from data protection agencies."* A few posts discussed "if a decline button is required," and the use of generators for cookie consents. ePD Art. 5 and GDPR Art. 7 were mentioned 16 and 14 times, respectively, while the EDPB's requirements for cookie consent were recommended only twice.

6.2.2 General Consent. Most posts in this category involve questions about when to ask for consent, what constitutes "lawful processing," and what falls under "legitimate interests". A poster asked about *"..a small business that has an application to save in-store credit for their clients. The only data being stored is the client's first and last name and how much in-store credit they have. [..] Do I need some written consent [..] to store their name?"* The responses were mixed and sometimes conflicting. Some argued that managing in-store credit is a "legitimate interests", while others noted that GDPR would apply as they "collect personal data". Other posts focused on when to ask explicit or implicit consent, what information should be included, and how to carry out age restrictions or obtain parental consent. A poster sought *"[..] our application is not DIRECTLY targeted to users <13 or <16, what would be the best action to go around this? Restrict users <13 [..]? and have <16 users requested parental consent?"* The responders mostly cited "The UK Information Commissioner's Office (ICO)" documentation or GDPR Art. 8.

Some posts were concerned about what to include in privacy policies, e.g., *"list all the 8 rights in a Privacy Policy"* or *"Does GDPR REQUIRES that companies publish detailed list (name, activities, location, etc) of all sub processors the company uses?"* Similarly, the responses were often contradictory: while the majority mentioned that they "do not have to list it *publicly* in the Privacy Policy", a few suggested to "give a list of where data will be shared". In these cases, it is up to the poster to follow up, refer to GDPR articles or guidelines, etc. to verify the responses. GDPR Art. 4, 6, and 14, as well as the ePD, CNIL, IAPP, EDPB, and ICO were cited.

6.2.3 DPA. The posts within this category looked for insights regarding the agreement between data processors and data controllers, what to include in a DPA document, or when they are required. One post asked *"if a DPA was required with their internet service provider for a self-hosted website, due to IP addresses counting as online identifiers."* Another post by a B2B cloud software provider asked *"if a DPA is necessary if their privacy policy and terms of service documents contain the same information."* Replies to this post sought clarification about the nature of the business, but all mentioned that a separate DPA document was not necessary. Some replies highlighted that while the business needs an agreement with their customers (the data controllers), no agreement was necessary with users of the customer's websites. One reply pointed out the source of confusion as being a distinction between the GDPR's definition of a DPA (a specific document) and the colloquial use of the term to refer to any agreement to meet the requirements of GDPR Art. 28.

6.2.4 Data Transfer. Posts in this category discussed the transfer of data to third countries, especially in the absence of any agreements. One post asked about *"access to sensitive data with remote workers in other countries."* The top response suggested that the poster conduct a risk assessment on the data and who currently has access. It also highlighted that additional steps would be needed if the remote workers are in non-EU countries. The original poster expressed confusion about whether there is a legal distinction between data at rest and data in transfer. Several other posts asked about how they could use US services as an EU-based company. For example, one post listed US-based transactional services and asked if they can be used in a compliant manner. Another asked if their EU-based SaaS, which has exclusively North American customers, needed to comply with GDPR. A response pointed to Art. 3, as well as the EDPB guidelines on the territorial scope of the GDPR. A significant portion of the discussion on these posts involved the Schrems 2 case, which invalidated the EU-US Privacy Shield. Responses often

conflicted, with responders pointing to information published both before and after this case, causing significant confusion.

6.2.5 DPIA.* Users also requested advice about DPIAs, including what these documents entail and how to assess risks prior to starting a project. The content of posts in this category varied significantly, but the general trend indicated that users were often unaware of when DPIAs were necessary. For example, a user planned to *"implement real-time tracking mechanisms"* and sought advice on *"how to clearly mitigate the risks prior to deployment."* One reply simply advised the poster to conduct a formal risk assessment before continuing. Another response agreed with this suggestion, but also discussed the implications of such tracking, highlighting GDPR Art. 5, 11, 13, and 17. The same response pointed to both the EDPB/WP29 and the ICO guidelines, *"even outside of the UK."* Another post discussed starting an E-learning forum and sought advice on logging practices prior to developing the software. The poster was uncertain whether they must log viewership details of each user, so that users may know who accessed their data. Responses to this post asked for further details about the goals of such logging, highlighting that there was no GDPR requirement for this use case. One reply specifically claimed that the opposite may be true, that *"such data collection could easily be a GDPR violation."*

Another post, by a user who stated they work at a privacy consulting firm, asked if a DPIA was necessary for the sensor data of a specific medical instrument that monitors patient movements, mentioning that their client is an "elderly care house". One response discussed their own related experience and broadly recommended that a DPIA be completed. Another asked for follow-up details, such as the country the poster was in, and referred to the Italian Supervisory Authority, which also recommends conducting a DPIA.

6.2.6 DSAR.* Concerns about DSARs revolved around situations where users request access to their data. One poster expressed confusion as to whether it includes more than just PII, asking *"When a user requests a copy of their information, is this all data, or only PII?"* The responses were conflicting, with one user pointing to Google as a baseline, suggesting that *"I think they have to give you everything."* Another answered that only the personal data the user directly provides must be given, but nothing derivative from it. Several posts asked broadly how other community members were managing DSARs or which tools they were using to make handling requests easier. One common response when posters were asking about data deletion was to highlight that *"the data deletion request right is not absolute"*, and that data retention is appropriate when there is a legal obligation to do so. One post by an employee of an educational platform for institutions asked about a deletion request they received directly from one of their client's users. The consensus among all responses was that it is the data controller's responsibility to handle such requests, not the data processor. Some replies linked to ICO guides for supporting information on individual rights and the obligations between controllers and processors.

6.2.7 RoPA.* A small portion of posts sought advice directly on RoPA documents, as well as what is considered PII requiring RoPA, platforms to manage RoPA, and on data retention requirements tied to RoPA. One user asked *"how to keep RoPA organized in a large organization"* and sought recommendations for *"more efficient management solutions."* Another asked *"whether a referrer URL, user agent, and three-quarters of an IP address together count as personal data and whether the anonymization process falls under RoPA?"* In addition to Art. 30, many of the responses referred to GDPR articles such as Art. 13, 14, 17, and 28. For example, one response discusses *""You will have mapped all of the personal data you process to the purpose and lawful basis for your RoPA (art 30) and that will align with your privacy notice (art 13-14). It should be fairly easy to add your retention to the above."* The discussions on RoPA and the number of articles to consider indicate its complexity.

6.2.8 Data Breach Notification.* A few posts discussed the implications of data breaches, how they should be handled, what constitutes a data breach under GDPR, what the responsibilities of the third-party are, and when to report the breach. For example, one user asked about *"Third-Party Servicer Compromised in Data Leak that (potentially) contained PII information on our clients, some of whom reside within UK/EU Citizen Information. Who discloses this to the GDPR Regulatory Authority? My company or the third-party servicer?"* The discussions revolved around the fact that it is the data controller's responsibility. Most posts had a follow-up question from the poster. In the case above, the poster also asked whether the number of affected people matters in this context. The response to what constitutes a data breach had some confusion, e.g., one responder mentioned *"as the definition of personal data breach in the law appears to exclude temporary loss of availability of data. However, guidance from the ICO and EDPB both appear to include it."*

6.2.9 GDPR Definitions & Generic Compliance. Some users needed clarification on basic GDPR terminologies or expressed confusion about ensuring GDPR compliance. Posts in these two categories did not closely relate to other major challenges. For example, many were unsure how to treat certain types of data under GDPR. One user asked about *"which pieces of medical information needed to be handled differently within their app."* Another sought clarification on the *"differences between a data processor and a data controller."* Regarding compliance, another post asked *"if GDPR applied to their business if they use a third-party payment processor and do not collect any data themselves."* Among these posts, responses often cited specific GDPR articles and EDPB guidelines for clarification.

6.2.10 Others. 9% of posts raise concerns not aligned with any of the challenge categories. These posts reflect diverse issues, like users promoting their own privacy tools or developers looking for resources to learn about privacy regulations. One post from an Australian developer *"expressed frustration about the perceived jurisdictional overreach of GDPR, questioning the EU's authority to regulate non-EU businesses and proposing to block EU users."*

7 Discussion

Key Insights and Implications. Our findings suggest that practitioners use these forums to navigate privacy compliance challenges in practice. Although our survey was posted in EU-focused subreddits, most participants were from North America. Despite 62% of survey respondents reporting they had in-house privacy counsel, ~38% either had counsel without privacy experience, no counsel, or were unsure. This dichotomy suggests several possibilities supported by a recent study of 51 US privacy attorneys [42]: in-house

self-service resources may be limited, do not exist, or are unsituated in design contexts; those with legal advice see opportunities to share their experiences with those who lack access to advice; and/or there is a translation challenge wherein access to legal advice does not guarantee that it can be easily or effectively translated into compliance design decisions. Our results indicate that while the presence of in-house legal teams increases confidence in companies' privacy practices [28, 48, 57], participants continue to engage with external forums even when internal expertise is available.

Our analyses show there is high motivation to provide advice (83%) and to share developer perspectives and experiences (82%). This high motivation creates challenges in determining what information is most credible. Respondents report a variety of means to check the credibility of responses, including checking poster profiles (22%), searching the web (18%), studying the post thread for citations (17%), among others. Overall, however, the means to assess credibility are ad hoc and rely heavily on features provided by Reddit (e.g., upvoting, threaded comments that may include critiques). ~11% report rarely checking the credibility. Despite the earlier result that most participants have in-house privacy counsel (62%), or legal counsel in general (87%), a key finding is the risk that developers may still encounter inaccurate or inconsistent information that is difficult to evaluate.

Finally, we found that discussions on regulatory-focused subreddits do cover permission-related issues, similar to prior work [38, 45, 62], and that consent, especially cookie consent, was the most prominent compliance challenge discussed. However, unlike prior work, we uncovered a broader set of challenges regarding consent, focusing mainly on design decisions, e.g., the need for a decline button, which design features to include, and when consent is required, which are motivated by nuances in shifting interpretations of privacy law. For example, the European Data Protection Board clarified in 2020 that declining cookies must be as easy as accepting cookies[1], and in 2023 clarified that the reject cookies button must be prominent displayed[2], over five years after the GDPR became effective. Despite clarifications from regulators, these challenges persist, highlighting the need for more nuanced practical guidance.

Future Research Directions. Developers' reliance on external advice may indicate a lack of in-house self-service resources for privacy-by-design (PbD) practices. If so, this creates an opportunity for the development of techniques that align with conventional legal advice, avoid dark patterns, and are grounded in design contexts. While regulators provide broad compliance guidance intended to address design questions, such guidance is often limited and insufficiently contextualized. Advances in LLMs have led to human-AI design ideation [51], including the exploration of design alternatives beyond code generation [70]. Investigating human-AI collaboration on PbD may support the need identified in our analysis. In addition, techniques to synthesize privacy requirements from laws could make regulatory information more accessible to developers, e.g., by extracting reusable requirements from privacy laws [4].

Another key finding is the difficulty that users face when evaluating privacy advice in subreddits. Prior research has examined how general Internet users judge the credibility of web pages [17] and search results [68], and evaluate health-related information [9]. However, these sources are often manipulated to monetize the public, whereas other issues underpin the credibility of privacy advice, such as the advisers' access to reliable information and its comprehension, the shift in legal interpretation and regulator guidance, and clarity of communication between the poster and reader. New research that reconciles posted claims with broader legal contexts or estimates poster credibility based on message history (e.g., how often they are consistent with established facts) could provide new tools for fact-checking PbD advice from forums.

Focusing directly on communication, it is essential to examine closely how posters and responders surface assumptions and facts in the posts and resolve misunderstandings and ambiguities during discussions [49]. Modeling these interaction patterns could support the design of dialogue policies (i.e., a set of rules to guide the models) for LLM-based privacy assistants and compliance tools [13, 14].

Educational Takeaways. Our findings show that users seek GDPR advice on forums, highlighting a gap in privacy compliance education and training, which is consistent with prior work [48]. We observe that users are unfamiliar with GDPR requirements and guidelines, such as the EDPB's Cookie Banner Report (1/18/23), which is cited infrequently. These results emphasize the need for *companies* and *educators* to create modules, courses, and training materials on topics such as consent, PII identification, data processing agreements, data transfers, and broader regulatory compliance. Our survey results show that having privacy-related experience positively impacts how users interact on forums and helps them better critically assess the credibility of responses; yet, a quarter of participants do not have internal corporate training. This gap illustrates how privacy professional trade groups, such as the International Association of Privacy Professionals, could contribute to ongoing professional learning for engineers on PbD topics.

Implications for Industry and Regulators. Accurately implementing cookie consent banners and obtaining consent are the top challenges identified in our findings. Thus, companies and open source developers could create Standard Development Kits (SDKs) with privacy-related UI components, such as cookie banners, alongside other reusable components that avoid dark patterns [1]. Automated tools that identify privacy-related source code and use machine learning to generate privacy notices could assist developers [34, 35] if they were made more easily accessible through IDE plugins. The need for tools and templates for creating effective DPA, RoPA, and DPIA was a recurring question. Regulators could provide templates and practical examples for such documents, while companies could develop tools to (partially) automate the process.

8 Conclusion

In this paper, we conducted a quantitative survey of 223 users on regulatory-focused subreddits and a qualitative analysis of 2,248 posts using an LLM-based approach to understand users' professional expertise, common privacy challenges, and motivations and practices for seeking advice or responding on Reddit. We provide detailed analysis of GDPR-related challenges, the interaction and engagement behavior of Reddit users, credibility assessment strategies, and how they impact decision-making processes. We discuss future research and education directions based on our findings.

[1] Guidelines 05/2020 on consent under Regulation 2016/679, Adopted on 4 May 2020
[2] The report undertaken by the Cookie Banner Taskforce, Adopted on 17 Jan. 2023

Acknowledgments

The authors used generative AI-based tools to correct typos, grammatical errors, and awkward phrasing.

This research was funded by NSF CAREER #2238047 and NSF Award #2217572.

References

[1] Sanju Ahuja, Johanna Gunawan, Nataliia Bielova, and Cristiana Teixeira Santos. 2026. Dark Patterns and the EU Digital Services Act: Mapping Autonomy Violations and Design Factors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–30.

[2] Noura Alomar and Serge Egelman. 2022. Developers Say the Darnedest Things: Privacy Compliance Processes Followed by Developers of Child-Directed Apps. *Proceedings on Privacy Enhancing Technologies* (2022).

[3] Arash Amini, Yigit Ege Bayiz, Ashwin Ram, Radu Marculescu, and Ufuk Topcu. 2025. News Source Credibility Assessment: A Reddit Case Study. In *Proceedings of the International AAAI Conference on Web and Social Media*. 68–82.

[4] Wilder Baldwin, Shashank Chintakuntla, Shreyah Parajuli, Ali Pourghasemi, Ryan Shanz, and Sepideh Ghanavati. 2025. Generating Privacy Stories from Software Documentation. In *Proceedings of the IEEE International Requirements Engineering Conference (RE)*. 432–440.

[5] Bryce Boe. 2026. PRAW: Python Reddit API Wrapper. https://pypi.org/project/praw/ Accessed: May 9, 2026.

[6] Norman Breslow. 1970. A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship. *Biometrika* 57, 3 (1970), 579–594.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[8] Centre for Information Policy Leadership. 2017. Top 10 Operational Impacts of the GDPR: Implementation Challenges and Strategies. https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/final_cipl_gdpr_implementation_challenges_summary_document_27_april_2017.pdf. Accessed February 10, 2025.

[9] Yung-Sheng Chang, Yan Zhang, and Jacek Gwizdka. 2021. The effects of information source and eHealth literacy on consumer health information credibility evaluation behavior. *Computers in Human Behavior* 115 (2021), 106629. https://doi.org/10.1016/j.chb.2020.106629

[10] Seung Youn Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. 2017. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement* 56, 10 (2017), 15–23.

[11] Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101.

[12] Asmita Dalela, Saverio Giallorenzo, Oksana Kulyk, Jacopo Mauro, and Elda Paja. 2021. A Mixed-Method Study on Security and Privacy Practices in Danish Companies. *arXiv preprint arXiv:2104.04030* (2021).

[13] Zachary Eberhart and Collin McMillan. 2021. Dialogue Management for Interactive API Search. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 274–285.

[14] Zachary Eberhart and Collin McMillan. 2022. Generating Clarifying Questions for Query Refinement in Source Code Search. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 140–151.

[15] European Union. 2024. The EU General Data Protection Regulation (GDPR). http://www.eugdpr.org/. Accessed February 10, 2025.

[16] Europrivacy. 2024. EIPACC – European Information Privacy and Cybersecurity Certification. https://www.europrivacy.org/eipacc. Accessed May 22, 2025.

[17] B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. How do users evaluate the credibility of Web sites? a study with over 2,500 participants. In *Proceedings of the 2003 Conference on Designing for User Experiences* (San Francisco, California) *(DUX '03)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/997078.997097

[18] Adrian Ford, Ameer Al-Nemrat, Seyed Ghorashi, and Julia Davidson. 2021. The Impact of Data Breach Announcements on Company Value in European Markets. In *WEIS 2021: The 20th Annual Workshop on the Economics of Information Security*. 1–8.

[19] Adrian Ford, Ameer Al-Nemrat, Seyed Ali Ghorashi, and Julia Davidson. 2023. The Impact of GDPR Infringement Fines on the Market Value of Firms. *Information & Computer Security* 31, 1 (2023), 51–64.

[20] Floyd J Fowler Jr. 2013. *Survey research methods*. Sage publications.

[21] Ron Garland et al. 1991. The Mid-Point on a Rating Scale: Is It Desirable? *Marketing Bulletin* 2, 1 (1991), 66–70.

[22] David Gefen, Shahin Jabbari, Rezvaneh Shadi Rezapour, Aria Pessianzadeh, Kshitij Kayastha, and Hilde Van den Bulck. 2024. The Evolving Meaning of Trust and Risk in Reddit Discourse About ChatGPT. *AMCIS Proceedings* (2024).

[23] Priscilla E Greenwood and Michael S Nikulin. 1996. *A Guide to Chi-Squared Testing*. John Wiley & Sons.

[24] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. 2018. Privacy by Designers: Software Developers' Privacy Mindset. *Journal of Empirical Software Engineering* (2018).

[25] Hamza Harkous, Sai Teja Peddinti, Rishabh Khandelwal, Animesh Srivastava, and Nina Taft. 2022. Hark: A Deep Learning System for Navigating Privacy Feedback at Scale. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2469–2486.

[26] Charles R Henderson. 1953. Estimation of variance and covariance components. *Biometrics* 9, 2 (1953), 226–252.

[27] Rashina Hoda. 2025. Qualitative Research with Socio-Technical Grounded Theory A Practical Guide to Qualitative Data Analysis and Theory Development in the Digital World. *Innovations* (2025).

[28] Stefan Albert Horstmann, Samuel Domiks, Marco Gutfleisch, Mindy Tran, Yasemin Acar, Veelasha Moonsamy, and Alena Naiakshina. 2024. "Those Things are Written by Lawyers, and Programmers are Reading That." Mapping the Communication Gap Between Software Developers and Privacy Experts. *Proceedings on Privacy Enhancing Technologies* (2024).

[29] Stefan Albert Horstmann, Sandy Hong, David Klein, Raphael Serafini, Martin Degeling, Martin Johns, Veelasha Moonsamy, and Alena Naiakshina. 2025. "Sorry for Bugging you so much." Exploring Developers' Behavior Towards Privacy-Compliant Implementation. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1215–1233.

[30] Stefan Albert Horstmann, Sandy Hong, Maziar Niazian, Cristiana Santos, and Alena Naiakshina. 2025. "I Need to Learn Better Searching Tactics for Privacy Policy Laws." Investigating Software Developers' Behavior When Using Sources on Privacy Issues. *arXiv preprint arXiv:2511.08059* (Nov. 2025). https://arxiv.org/abs/2511.08059v1

[31] Hugging Face. 2025. Hugging Face – The AI Community Building The Future. https://huggingface.co/. Accessed: November 13, 2025.

[32] International Association of Privacy Professionals. 2024. IAPP Certifications. https://iapp.org/certify/. Accessed: June 1, 2024.

[33] Leonardo Horn Iwaya, Muhammad Ali Babar, and Awais Rashid. 2023. Privacy Engineering in the Wild: Understanding the Practitioners' Mindset, Organisational Aspects, and Current Practices. *IEEE Transactions on Software Engineering* (2023).

[34] Vijayanta Jain, Sepideh Ghanavati, Sai Teja Peddinti, and Collin McMillan. 2023. Towards Fine-Grained Localization of Privacy Behaviors. In *IEEE 8th European Symposium on Security and Privacy*. 258–277. https://doi.org/10.1109/EuroSP57164.2023.00024

[35] Vijayanta Jain, Sepideh Ghanavati, Sai Teja Peddinti, and Collin McMillan. 2026. Automated Generation of Accurate Privacy Captions From Android Source Code Using Large Language Models. *arXiv preprint arXiv:2601.06276* (2026).

[36] Vijayanta Jain, Sanonda Datta Gupta, Sepideh Ghanavati, Sai Teja Peddinti, and Collin McMillan. 2022. PAcT: Detecting and Classifying Privacy Behavior of Android Applications. In *Proc. of the 15th ACM Conf. on Security and Privacy in Wireless and Mobile Networks*. ACM, 104–118.

[37] Tianshi Li, Yuvraj Agarwal, and Jason I Hong. 2018. Coconut: An IDE Plugin for Developing Privacy-Friendly Apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–35.

[38] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I Hong. 2021. How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of A Developer Forum on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[39] Jenny T Liang, Chenyang Yang, and Brad A Myers. 2024. A Large-Scale Survey on The Usability of AI Programming Assistants: Successes and Challenges. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–13.

[40] Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review* 57 (2024), 265–292.

[41] Jai Kruthunz Naveen Kumar, Aishwarya Umeshkumar Surani, Harkirat Singh, and Sanchari Das. 2026. Privacy Discourse and Emotional Dynamics in Mental Health Information Interaction on Reddit. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*. Association for Computing Machinery, 491–496.

[42] Thomas B. Norton and Travis D. Breaux. 2026. Privacy by Design's Last Mile: Practitioner Evidence of Translation Friction at the Legal-Engineering Interface. *In Submission: International Data Privacy Law* (2026).

[43] OpenAI. 2025. API Platform. https://openai.com/api/. Accessed: November 13, 2025.

[44] Nicholas Pangakis and Samuel Wolken. 2024. Keeping Humans in the Loop: Human-Centered Automated Annotation with Generative AI. *arXiv preprint arXiv:2409.09467* (2024).

[45] Jonathan Parsons, Michael Schrider, Oyebanjo Ogunlela, and Sepideh Ghanavati. 2023. Understanding Developers Privacy Concerns Through Reddit Thread

Analysis. *Joint Proc. of REFSQ-2023 Workshops, Doctoral Symposium, Posters & Tools Track and Journal Early Feedback co-located with the 28th Int. Conf. on Requirements Engineering: Foundation for Software Quality (REFSQ 2023), Barcelona, Catalunya* (2023).

[46] PECB. 2024. Certified Data Protection Officer (CDPO). https://pecb.com/en/education-and-certification-for-individuals/gdpr/certified-data-protection-officer. Accessed: May 22, 2025.

[47] Aria Pessianzadeh, Naima Sultana, Hildegarde Van den Bulck, David Gefen, Shahin Jabbari, and Rezvaneh Rezapour. 2025. In Generative AI We (Dis) Trust? Computational Analysis of Trust and Distrust in Reddit Discussions. *arXiv preprint arXiv:2510.16173* (2025).

[48] Maxwell Prybylo, Sara Haghighi, Sai Teja Peddinti, and Sepideh Ghanavati. 2024. Evaluating Privacy Perceptions, Experience, and Behavior of Software Development Teams. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*. USENIX Association, 101–120.

[49] Sarah Santos, Sara Haghighi, Sepideh Ghanavati, Travis D Breaux, and Thomas B Norton. 2024. Patterns of Inquiry in a Community Forum for Legal Compliance with Privacy Law. In *2024 IEEE 32nd International Requirements Engineering Conference Workshops (REW)*. IEEE, 251–259.

[50] Raphael Serafini, Stefan Albert Horstmann, and Alena Naiakshina. 2024. Engaging Company Developers in Security Research Studies: A Comprehensive Literature Review and Quantitative Survey. In *33rd USENIX Security Symposium (USENIX Security 24)*. 3277–3294.

[51] Hanshu Shen, Lyukesheng Shen, Wenqi Wu, and Kejun Zhang. 2025. IdeationWeb: Tracking the Evolution of Design Ideas in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan) *(CHI'25)*. Association for Computing Machinery, New York, NY, USA, Article 146, 19 pages. https://doi.org/10.1145/3706598.3713375

[52] Daniel J Solove. 2005. A Taxonomy of Privacy. *U. Pa. l. Rev.* 154 (2005), 477.

[53] Georgios Spanos and Lefteris Angelis. 2016. The Impact of Information Security Events to the Stock Market: A Systematic Literature Review. *Computers & Security* (2016), 216–229.

[54] Charles Spearman. 1961. The Proof and Measurement of Association Between Two Things. (1961).

[55] Mohammad Tahaei, Julia Bernd, and Awais Rashid. 2022. Privacy, permissions, and the health app ecosystem: A stack overflow exploration. In *Proceedings of the 2022 European Symposium on Usable Security*. 117–130.

[56] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. 2021. Deciding on personalized ads: Nudging developers about user privacy. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 573–596.

[57] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. 2021. Privacy champions in software teams: Understanding their motivations, strategies, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[58] Mohammad Tahaei, Tianshi Li, and Kami Vaniea. 2022. Understanding Privacy-Related Advice on Stack Overflow. *Proceedings on Privacy Enhancing Technologies* 2022, 2 (2022), 114–131. https://doi.org/doi:10.2478/popets-2022-0038

[59] Mohammad Tahaei, Marvin Ramokapane, Tianshi Li, Jason I Hong, and Awais Rashid. 2022. Charting App Developers' Journey through Privacy Regulation Features in Ad Networks. In *The 22nd Privacy Enhancing Technologies Symposium*. 33–56.

[60] Mohammad Tahaei and Kami Vaniea. 2019. A survey on developer-centred security. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 129–138.

[61] Mohammad Tahaei, Kami Vaniea, and Awais Rashid. 2022. Embedding Privacy into Design through Software Developers: Challenges and Solutions. *IEEE Security & Privacy* (2022), 49–57.

[62] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. 2020. Understanding Privacy-Related Questions on Stack Overflow. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[63] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The Science of Detecting LLM-Generated Text. *Commun. ACM* (2024).

[64] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (1996), 267–288.

[65] Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. (2024). arXiv:2402.05129 [cs.CL] https://arxiv.org/abs/2402.05129

[66] Yang Wang. 2009. Privacy-Enhancing Technologies. In *Handbook of research on social and organizational liabilities in information security*. IGI Global, 203–227.

[67] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems* (2022), 24824–24837.

[68] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1235–1244. https://doi.org/10.1145/1978942.1979126

[69] Mohammad Amin Zadenoori, Liping Zhao, Waad Alhoshan, and Alessio Ferrari. 2025. Automatic Prompt Engineering: The Case of Requirements Classification. In *Requirements Engineering: Foundation for Software Quality: 31st International Working Conference, REFSQ 2025, Barcelona, Spain, April 7–10, 2025, Proceedings*. Springer-Verlag, 217–225.

[70] J.D. Zamfirescu-Pereira, Eunice Jun, Michael Terry, Qian Yang, and Bjoern Hartmann. 2025. Beyond Code Generation: LLM-supported Exploration of the Program Design Space. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 153, 17 pages. https://doi.org/10.1145/3706598.3714154

A Data Availability

The repository containing the list of survey questions and the Reddit post analysis, including the ground truth and labeled datasets, all prompts, scripts, sample data for prompts, and results, can be found here: https://github.com/PERC-Lab/GDPR_discussion_data.

B Research Hypotheses and Results

B.1 Variability in Motivations for Participating in GDPR-related Discussions

- **H1a:** Users' motivations for participating in GDPR-related discussions vary by job roles.
- **H1b:** Users' motivations for participating in GDPR-related discussions vary by privacy-related certifications.
- **H1c:** Users' motivations for participating in GDPR-related discussions vary by privacy-related experience.
- **H1d:** Users' motivations for participating in GDPR-related discussions vary by company size.

Table 8 summarizes the statistically significant results of the hypothesis analysis.

B.2 Variability in GDPR Challenges

- **H2a:** The selection of specific GDPR challenges varies based on company type.
- **H2b:** The selection of specific GDPR challenges varies based on company size.
- **H2c:** The selection of specific GDPR challenges varies based on company location.
- **H2d:** The selection of specific GDPR challenges varies based on participants' job roles.
- **H2e:** The selection of specific GDPR challenges varies based on participants' privacy-related certifications.

Table 9 shows the results of the hypotheses analysis.

Table 9: Results for Hypothesis H2a to H2e

	Company			Job Role	Pri-Certified
	Type	Size	Location		
Statistic	$\chi^2(24) = 59.3299$	$\chi^2(12) = 24.1388$	$\chi^2(36) = 44.0361$	$\chi^2(28) = 27.2474$	$\chi^2(4) = 26.1499$
p-value	0.0001	0.0195	0.1681	0.5048	< 0.001
Cramér's V	0.2123	0.1564	0.1832	0.1439	0.2819
Sample Size	190	190	189	190	190

B.3 Influence of Legal Support on Compliance Confidence

- **H3a:** Organizations with in-house privacy legal teams report higher confidence in privacy compliance.
- **H3b:** Organizations that frequently seek external legal advice report lower confidence in privacy compliance.

Table 8: Statistically significant results for Hypothesis H1a to H1d

Factor	Motivation	Statistic	p-value	Effect Size	Sample Size
Privacy Certifications	Create posts to share knowledge/experience	$U = 956.50$	0.001	$r = 0.476$	N = 94
Privacy Certifications	Create posts to express opinions	$U = 859.50$	0.032	$r = 0.310$	N = 95
Privacy-related Experience	Create posts to share knowledge/experience	$H = 14.30$	0.006	$\epsilon^2 = 0.112$	N = 95
Privacy-related Experience	Reply to share perspectives/experience	$H = 8.80$	0.032	$\epsilon^2 = 0.051$	N = 117
Company Size	Create posts to express opinions	$H = 12.21$	0.016	$\epsilon^2 = 0.328$	N = 30
Company Size	Create posts to seek advice/feedback	$H = 10.96$	0.027	$\epsilon^2 = 0.279$	N = 30
Company Size	Reply to share perspectives/experience	$H = 9.02$	0.029	$\epsilon^2 = 0.140$	N = 47

Table 10 shows the results of the hypotheses analysis.

Table 10: Results for Hypothesis H3a and H3b

	In-house Legal Team	External Legal Advice
Statistic	$H = 8.4773$	$\rho = 0.0083$
P-values	0.0371	0.9096
Effect Size	$\eta^2 = 0.0294$	Negligible
Sample Size	190	191

B.4 GDPR Knowledge Sources

- **H4a:** Participants' primary sources of GDPR knowledge vary by job roles.
- **H4b:** Participants' primary sources of GDPR knowledge vary by company size.
- **H4c:** Participants' primary sources of GDPR knowledge vary by educational background.
- **H4d:** Participants' primary sources of GDPR knowledge vary based on privacy-related certifications.

Table 11 shows the results of the hypotheses analysis.

Table 11: Result for Hypothesis H4a and H4d

	Job Role	Company Size	Education	Pri-Certified
Statistic	$\chi^2(56) = 31.24$	$\chi^2(32) = 18.58$	$\chi^2(32) = 15.09$	$\chi^2(8) = 22.19$
p-value	0.9970	0.9717	0.9951	0.0046
Cramér's V	0.0830	0.1290	0.1163	0.2820
Sample Size	191	68	68	191

B.5 Credibility Evaluation by Professional Background

- **H5a:** Participants' credibility evaluation strategies are associated with their job role.
- **H5b:** Participants' credibility evaluation strategies are associated with their educational background.
- **H5c:** Participants' credibility evaluation strategies are associated with their company size.
- **H5d:** Participants' credibility evaluation strategies are associated with their privacy-related certifications.

Table 12 shows the results of the hypotheses analysis.

C Participants' Demographic Information

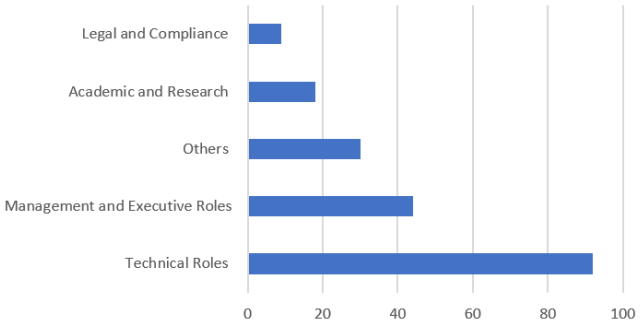
Table 13 shows the demographic information of the participants.

Table 12: Results for Hypothesis H5a to H5d

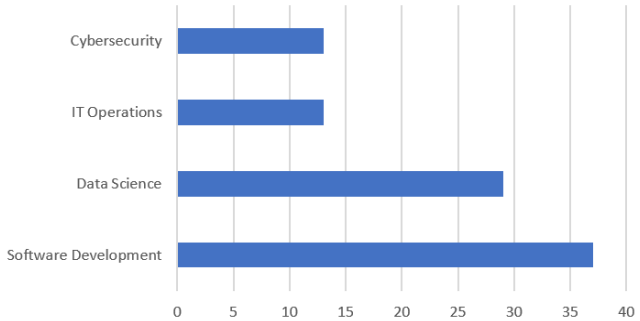
	Job Role	Company Size	Education	Pri-Certified
Main Effect (Group)	$F(7, 86) = 2.12$	$F(4, 24) = 2.00$	$F(4, 24) = 0.52$	$F(1, 92) = 0.57$
p-value (Group)	0.0498	0.1260	0.7186	0.4521
η_p^2 (Group)	0.147	0.250	0.080	0.006
Interaction	$F(42, 516) = 1.27$	$F(24, 144) = 2.43$	$F(24, 144) = 0.60$	$F(6, 552) = 0.80$
p-value (Interaction)	0.1270	<0.001	0.9251	0.5692
η_p^2 (Interaction)	0.093	0.288	0.091	0.009
Sample Size	94	29	29	94

D Distribution of Participants' Roles

Figure 5a and Figure 5b show the distribution of job roles and the subcategories within technical roles.



(a) Distribution of the Job Roles



(b) Top Four Technical Roles Subcategories

Figure 5: Distribution of Participants' Roles

Table 13: Demographic Information about the Participants

Category	Details
Gender	Female (26%), Male (73%), Non-binary (1%)
Age	18-24 (8%), 25-34 (40%), 35-44 (46%), 45-64 (6%)
Education	High School (3%), Associate (4%), Undergraduate (21%), Graduate (34%), Professional (39%)
Degree	CS (35%), Data Science (22%), IT (21%), Business (19%), Other (6%)
Company Size	1-5 (4%), 6-20 (16%), 21-50 (28%), 51-100 (23%), 101+ (30%)

E Distribution of Privacy Certification Types

Figure 6 shows privacy-related certifications dominate the responses.

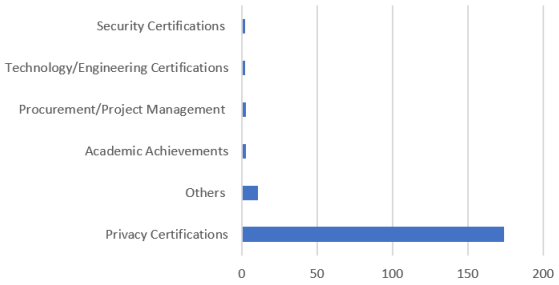


Figure 6: Categories of Privacy Certification Types

F Frequency of GDPR-related Engagements

Figure 7 shows the GDPR-related engagements.

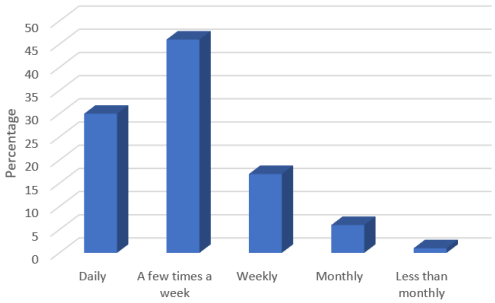


Figure 7: Frequency of GDPR-related Engagements

G Frequency of Posting or Replying

Figure 8 shows the frequency of users' posting and replying activity on GDPR-related Reddit posts during the last six months.

H Frequencies of Interaction Types on Reddit

Table 14 shows the frequency of users' interaction types on Reddit.

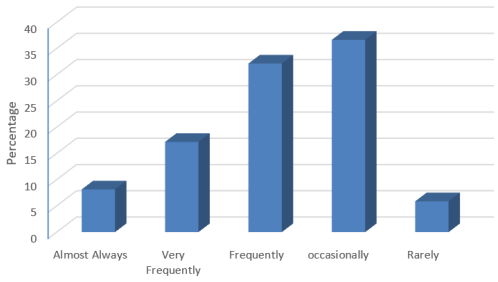


Figure 8: Frequency of GDPR-related Posting/Replying

Table 14: Frequencies of Interaction Types on Reddit

Interaction Type (Q17)	% (Count)
Primarily create original posts	19% (41)
Mostly reply to other users' posts	31% (68)
Equally create original posts and reply	34% (75)
Mostly read posts w/o posting or replying	17% (37)

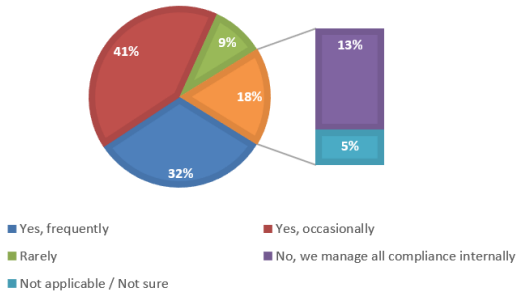


Figure 9: Use of External Privacy or Legal Advice

I Use of External Privacy or Legal Advice

Figure 9 shows the distribution of participants' use of external privacy or legal advice.

J Shared GDPR Challenges in Survey vs. Posts

Figure 10 shows the proportion of GDPR challenges present in both the survey and Reddit post analysis.

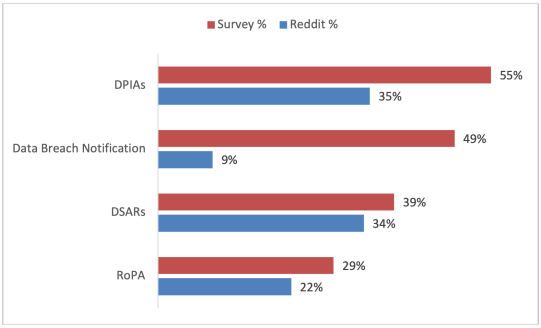


Figure 10: Shared GDPR Challenges in Survey vs. Posts

K Reddit Posts Analysis

K.1 Privacy Terms for Reddit Queries

Table 15 shows the terms extracted from the Hark methodology and their expanded synonyms. Some terms were removed from the final set, but since their synonyms were retained, we include them.

Table 15: Privacy-related Terms used for Reddit Queries

Harks Terms [25]	Added Synonym Terms
Privacy	—
Surveillance	Monitoring, Tracking
Aggregation	Collection, Use, Processing, Sharing, Transfer
Identification	Authentication, Profiling
~~Secondary Use~~	Data Repurposing
Exclusion	—
Breach	Vulnerability, Compromise, Unauthorized, Access, Exposure, Leak
Confidentiality	Anonymity
Disclosure	—
Notice	—
Data Minimization	Data Protection
Purpose Specification	—
~~Use Limitation~~	Access Control
Choice	—
Consent	Permission, Authorization, Agreement, Opt-in, Opt-out, Retention

K.2 Ground Truth Creation

Table 16 shows the results for each round of annotation when identifying posts related to software development and privacy.

K.3 Prompt Strategies and Results

Privacy and Development Labels. We developed parts of our CoT prompts following the guidelines described in [65, 69]. We accessed open-source models via Hugging Face [31] and OpenAI models through its API [43]. For all models except GPT-5 and GPT-5-mini, we set *temperature* = 0 and *top-p* = 1.0. For GPT-5 and GPT-5-mini, we set the verbosity and reasoning effort to "medium," and $top - p = 1.0$.

Listing 2 shows the CoT prompt used for detecting privacy and development posts. Note that the prompts for zero-shot and few-shot are shown on our GitHub repository.

Table 16: Annotation Summary for Relevant Posts

Rounds	Posts Reviewed	Disagreements	Cohen's Kappa	Relevant Posts
Round 1	100	11	0.749	32
Round 2	100	14	0.684	32
Round 3	100	9	0.808	34
Round 4	50	4	0.834	18
Total	**350**			**116**

Listing 2: CoT Prompt for Privacy & Development Posts

```
You are an expert system that needs to determine if Reddit posts
      are related to {topic} with either True or False and to
      provide an explanation for your reasoning.
Let's analyze how to classify posts step by step:
Step 1: Review the examples of posts and their answers.
Example 1: {question}
Post: {example 1}
Explanation: {reasoning for example 1}
Answer: {label of example 1}
Example 2: {question}
Post: {example 2}
Explanation: {reasoning for example 2}
Answer: {label of example 2}

Step 2: Understand how to apply this reasoning to new posts, using
       the examples and explanations as guidance.

Step 3: Apply this understanding to the following question and
      post.

Step 4: Determine the answer to this question for the following
      post. Make sure to review your response and provide a
      rationale for your selection:
{question}
Post: {post text}
Explanation:
```

We instantiated {question} from the two queries in Listing 3.

Listing 3: Questions Used in the Prompt for Privacy & Development Detection

```
Privacy Question: Is this Reddit post related to privacy or data
      protection issues under any context, including personal data,
       GDPR, or online privacy? Posts that are surveys are
      completely unrelated. Answer only with True or False.

Development Question: Is this Reddit post related to software,
      software development, developers, programming, engineering
      practices, software development activities, or code-related
      issues? Software development includes requirements analysis,
      software specification, design and development, quality
      assurance, testing, and maintenance. Posts that are surveys
      are completely unrelated. Answer only with True or False.
```

Table 17 compares the performance of various models in terms of accuracy, precision, recall, and F1 score on detecting privacy and development related posts using CoT prompting on test dataset.

Table 17: Performance on Detecting Privacy and Development Related Posts with Chain-of-Thought Prompting on Test Set

Model	Topic	Accuracy	Precision	Recall	F1 Score
Llama3.2-1B-Instruct	Privacy	0.7500	**0.9811**	0.7591	0.8560
	Development	0.4286	0.3034	0.6000	0.4030
Llama3.2-3B-Instruct	Privacy	0.7071	0.9800	0.7153	0.8270
	Development	0.6071	0.4000	0.4444	0.4211
Llama3.1-8B-Instruct	Privacy	0.8500	0.9754	0.8686	0.9189
	Development	0.7571	0.6897	0.4444	0.5405
Qwen2.5-7B-Instruct	Privacy	0.7714	0.9730	0.7883	0.8710
	Development	0.7643	**0.9286**	0.2889	0.4407
GPT-3.5-turbo-1106	Privacy	0.9429	0.9568	0.9429	0.9498
	Development	0.7429	0.7574	0.7429	0.6477
gpt-4o-mini	Privacy	**0.9500**	0.9570	**0.9500**	**0.9535**
	Development	0.4714	0.6592	0.4714	0.4552
gpt-5-mini	Privacy	0.9429	0.9568	0.9429	0.9498
	Development	0.8071	0.8022	0.8071	0.8008
gpt-5	Privacy	0.9286	0.9565	0.9286	**0.9423**
	Development	**0.8571**	0.8556	**0.8571**	**0.8531**

GDPR Challenges Labels. We randomly sampled 84 posts (i.e., development set) from the subset labeled with challenges to refine our prompt (i.e., zero-shot, one-shot, and CoT). The prompts were designed to yield only the most relevant challenge in each post to improve reliability. A generated label was counted as a true positive if it matched at least one label assigned to the post in the ground truth dataset; if the label did not match, then it was counted as a false positive; and if no label was generated, then it was counted as a false negative. Similar to the first task, we aim to maximize recall and minimize false negatives, since correcting labels is an easier task than identifying missing labels [36].

We instantiated labels' list from Table 20 and provided their corresponding definitions. This list covers all categories except "data transfer" and "generic compliance," which were introduced later after the LLM had classified a substantial portion of data under the category "Others."

Table 18 shows the results of detecting GDPR challenges of posts with CoT Prompting on the test set.

Table 19 shows the frequency of challenges labeled by GPT-5-mini, the number of false positive labels, and the final count after the manual review.

L GDPR Challenge Categories and Definitions

Table 20 defines the GDPR-related challenge categories used in the annotation process, including definitions, relevant legal basis, examples, and clarification notes of each category.

Table 18: Performance on Detecting GDPR-related Challenges using Chain-of-Thought Prompting on Test Set

Model	Accuracy	Precision	Recall	F1 Score
Llama3.2-1B-Instruct	0.3387	0.5527	0.3750	0.3584
Llama3.2-3B-Instruct	0.3387	0.4159	0.3438	0.2939
Llama3.1-8B-Instruct	0.4677	0.5353	0.4844	0.4768
Qwen2.5-7B-Instruct	0.6452	0.6772	0.6719	0.6388
GPT-3.5-turbo-1106	0.4677	0.4193	0.4688	0.4019
gpt-4o-mini	0.6613	0.7513	0.6875	0.6683
gpt-5-mini	**0.7258**	**0.7811**	**0.7377**	**0.7357**
gpt-5	0.6774	0.7149	0.6875	0.6857

Table 19: Frequency of Labeling Challenges with GPT-5-mini and Final Frequency after Review

Challenge	LLM Predictions	False Positive	Final Count
Others	51	6	133
Cookie-Consent	200	19	108
General-Consent			88
DPA	94	35	70
Definition	82	50	38
DPIA	22	10	22
DSAR	30	6	28
RoPA	21	7	17
Breach-notification	12	5	8
Total	512	138	512

Table 20: Codebook for GDPR Challenge Categories with Definitions and Examples

Category	Definition	Legal Basis	Examples	Source
Data Protection Impact Assessments (DPIAs)	Covers posts about Data Protection Impact Assessments (DPIAs). This is a risk assessment done before starting a project by the data controller that may affect people's privacy. It helps identify and mitigate potential risks associated with the use and processing of large volumes of personal data. Note that, DPIA is a privacy impact assessment document.	Article 35	"What are the risks of X?" "Is it safe to use X if you're an EU company?"	Survey
Data Breach Notification	Covers posts about data breach incidents. When personal data/information or PII is lost, stolen, or exposed, a breach occurs. Thus, the organization or data controller must notify data protection authorities within 72 hours of discovering a breach involving personal data.	Article 33	"Was this incident a breach?" "As a third party, if I were aware of a breach, should I report it?"	Survey
Data Subject Access Requests (DSARs)	Covers posts about Data Subject Access Requests (DSARs). These allow individuals to request access to their personal data collected by the organization. The organization or a data controller must reply within a short time and include all relevant information. The DSAR also includes other rights as follows: the right to erasure (right to be forgotten), the right to access by the data subject, the right to rectification, and the right to object.	Articles 15–17 and 21–22	"We received a data subject request asking us to X." "What information should be included in a DSAR response?" "Can I request to delete all my emails after leaving the company?"	Survey
Records of Processing Activities (RoPA)	Covers posts about Records of Processing Activities (RoPA). Based on RoPA, organizations or data controllers must keep detailed records of what personal data they process, for what purpose, how it's stored, and who it's shared with. This document is completed after the data collection and processing is done.	Article 30	"Does collecting X mean we must keep processing records?"	Survey
Consent	Covers posts that discuss whether it is required to obtain explicit or implicit consent or permission before collecting or processing personal data. Consent can be of type privacy policies, privacy notices, cookie notices or cookie banners. Parental consent, cookie consent, etc, are also included. Note that consent is not about data subject access rights/requests or other GDPR compliance.	Articles 6–8	"Do I need consent to do X?" "How do I ask for user consent before processing data?"	During Ground Truth Creation
Definitions	Covers posts that ask for clarification or explanations of terminology. Definition refers to defining what it means by personal information, data processor, data controller, what is PII, etc.	Article 4	"Is an IP address considered personal data under GDPR?" "Is a customer ID pseudonymised data under GDPR?"	During Ground Truth Creation
Data Processing Agreements (DPAs)	Covers posts about Data Processing Agreements (DPA) which is a legally binding contract between the data controller (e.g., company) and the data processor (e.g., a third party) for processing data, duration of the processing, the nature and purpose of the processing, the type of personal data and categories of data subjects and the obligations and rights of the controller. For example, when the data processor is an email client, a cloud storage service, or website analytics software, you must have a data processing agreement with each of these services. Note that DPA is an agreement document.	Article 28(3)	"Do I need a DPA with my cloud provider?" "What should be included in a DPA?"	During Ground Truth Creation
Data Transfer	Covers posts about transferring personal data to third countries (especially outside the EU) or to international organizations.	Articles 44-50	"How can I send PII to a US-based server?" "Does Google Font send users' PII outside the EU?"	After Full Dataset Labeled
Generic Compliance	Posts related to concerns surrounding simply complying with the GDPR, but lacking any specific relation to the other challenges.		"Does it comply with GDPR?"	After Full Dataset Labeled
Others	Posts that do not fit any of the above categories.		"Why does the GDPR apply to my company if it is established outside the EU?" "Are EU regulators allowed to conduct spot checks without prior notice?" "Does GDPR affect VPN usage under Swedish law?" "We created a privacy preserving tool."	After Full Dataset Labeled