Personal Data Flows and Privacy Policy Traceability in Third-party LLM Apps in the GPT Ecosystem

Juan-Carlos Carrillo
VRAIN, Universitat Politècnica de València
Spain
juaca10j@upv.es

Jose Luis Martin-Navarro
Aalto University - VRAIN, Universitat Politècnica de València
Finland - Spain
jose.martinnavarro@aalto.fi

Rongjun Ma
Aalto University
Finland
rongjun.ma@aalto.fi

Jose Such
INGENIO (CSIC-Universitat Politècnica de València)
Spain
jose.such@csic.es

Abstract

The rapid growth of platforms for customizing Large Language Models (LLMs), such as OpenAI's GPTs, has raised new privacy and security concerns, particularly related to the exposure of user data via third-party API integrations in LLM apps.To assess privacy risks and data practices, we conducted a large-scale analysis of OpenAI's GPTs ecosystem. Through the analysis of 5,286 GPTs and the 44,102 parameters they use through API calls to external services, we systematically investigated the types of user data collected, as well as the completeness and discrepancies between actual data flows and GPTs stated privacy policies. Our results highlight that approximately 35% of API parameters enable the sharing of sensitive or personally identifiable information, yet only 15% of corresponding privacy policies provide complete disclosure. By quantifying these discrepancies, our study exposes critical privacy risks and underscores the need for stronger oversight and support tools in LLM-based application development. Furthermore, we uncover widespread problematic practices among GPT creators, such as missing or inaccurate privacy policies and a misunderstanding of their privacy responsibilities. Building on these insights, we propose design recommendations that include actionable measurements to improve transparency and informed consent, enhance creator responsibility, and strengthen regulation.

Keywords

LLM apps, privacy policy, measurement

1 Introduction

The rapid growth of Large Language Models (LLMs) has led to the emergence of open ecosystems for creating and sharing LLM-driven applications. Platforms such as Poe [51], FlowGPT [71], Yuanqi [72] and OpenAI's GPT Store [46] enable users to easily build and distribute their own applications with minimal effort. Starting from a simple prompt, users can develop specialized LLM apps, for example, by uploading a math textbook and instructing the model to act as a tutor. While these ecosystems foster personalization and community-driven creativity exchange, they also introduce heightened security and privacy risks that have yet to be understood.

LLM app platforms are exposed to a range of security and privacy risks. Since these applications share the same underlying LLM backend, they inherit many of the vulnerabilities affecting general-purpose language models. These risks include unauthorized data processing, data leakage [29], prompt injection [20], training data extraction [10], backdoor attacks, and the profiling of individuals without consent [30]. Additionally, users often have little control or visibility over how their data is collected, reused, and exposed, which heightens the risk of privacy violations and misinformation dissemination [30, 76, 82, 83].

Due to the low barrier to entry for LLM app creation, the distinction between LLM app creators and end users is often blurred, leading to miscommunication and uncertainty about how user data is handled [35]. Unlike traditional application ecosystems, where developers are typically responsible for data protection and are expected to possess technical expertise [21], many creators in the LLM app ecosystem may lack the necessary experience to implement adequate security measures or legal safeguards for user data [35]. This raises important questions about how data handling is actually implemented in practice within these LLM app ecosystems.

To make it even more challenging, LLM app platforms increasingly enable integrations with third-party services to extend LLM app functionality. For example, GPTs hosted by OpenAI allow LLM apps to connect with external services [45]. This complicates data flows and creates challenges in securing inputs, outputs, and interactions across multiple components [7, 52]. These third-party integrations may also introduce external risks, including the possibility of unauthorized access through third-party APIs [78].

In this paper, we examine two key aspects: 1) we identify what types of personal information are transmitted by GPTs to external services (beyond OpenAI's infrastructure); and 2) we assess whether these data practices align with the transparency principles stated in the GPTs' privacy policies.

This paper makes several contributions. First, we develop and validate a scalable methodology to analyze personal data flows

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–23

https://doi.org/XXXXXXX.XXXXXXX

within GPTs, automatically identifying potential personal information (PI) transmissions to third-party APIs using a refined taxonomy and advanced classification techniques. Second, by applying this method to a large dataset of GPTs, we identify specific categories of PI that are frequently transmitted via external API calls. Third, we conduct a systematic assessment of the alignment between detected PI flows and the disclosures in the associated privacy policies, evaluating both traceability and completeness. We explicitly scope our analysis to the collection of data transmitted through the Action interface. Retention, downstream use, and cookie- or session-level tracking are excluded from this study. Based on these findings, we offer actionable recommendations to foster a more responsible and privacy-conscious LLM app platforms.

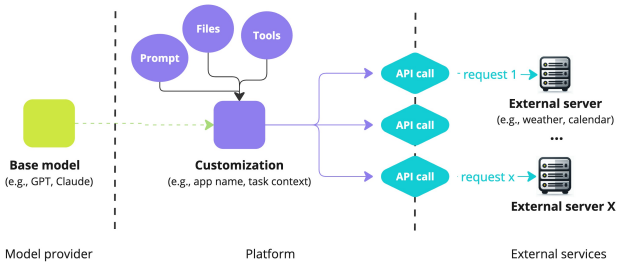

Figure 1: A common infrastructure for customized LLM applications, supporting both in-platform customization and integration with external services

2 Background

Large Language Model (LLM) apps are customizations of foundation models designed to perform specific tasks across a wide range of domains. For example, an LLM can be customized to function as a calendar assistant by integrating with services like Google Calendar to schedule appointments and deliver timely reminders for upcoming meetings and events [25, 59]. As opposed to general-purpose LLMs, these LLM apps empower users to become creators, allowing them to customize the model through prompt engineering, incorporate domain-specific knowledge for fine-tuning, or integrate external services [46]. To support the creation, discovery, and deployment of such LLM apps, various platforms have emerged, such as OpenAI's GPT Store. These platforms enable creators to publish their LLM apps and share them with a broader user community.

Most LLM app platforms provide prompt-based LLM app creation, where users define LLM app behavior through natural language prompts, such as Miniapp [41] and FlowGPT [71]. In this case any data disclosed by the user would be kept within the platform itself. In addition, there are other platforms that allow extending the functionality of LLM apps through API integration, enabling connections to third-party external services; examples include OpenAI's GPTs [46] and Poe [51]. This second type of platforms is the focus of our work. In particular, Figure 1 illustrates this type of architecture, where customized LLM apps can access external services exposed as API calls.

To illustrate the practical implications of such integrations and how privacy violations can occur, consider a calendar assistant GPT. The creator of this assistant might instruct it to capture a user's location, preferred times, and email address, and then forward these details to a third-party scheduling service. Even if the scheduling platform's own privacy policy claims it does not collect or store precise user location, the GPT developer could still transmit location data contradicting the stated practices of the service.

In this paper, we focus on OpenAI's GPTs store, which has gained significant popularity and supports integration with third-party services. All GPTs (LLM apps) run on OpenAI's hosted infrastructure and can be kept private, shared within a team, or get published to the public GPT Store [46]. As of January 2024, over three million GPTs had been created, attracting approximately 6.1 million monthly visits [47, 62]. OpenAI's GPTs allow users to build their own LLM apps on top of GPT-4o [24] without writing code. Users can tailor the GPTs' behavior, tone, and responses by providing instructions, examples, and optional reference files. Creators of GPTs may enable access to internal tools of OpenAI's infrastructure, such as a code interpreter or OpenAI's image generation model DALL-E. Additionally and very importantly, **GPTs may connect to external services through features called "Actions"**. Creators must define each "Action", specifying the parameters to be sent during execution and providing a link to a privacy policy. The action description and its parameters are passed to the GPT model as part of the GPT context. When a user interacts with the GPT, the model interprets if it needs to use the action and extracts the required parameters from the user prompt. That is, the model extracts from the prompt what it interprets fits the definition given by the GPT creator.

3 Methodology

Our overarching goal is to measure the extent to which third-party GPTs collect personal information and whether their privacy policies accurately reflect these practices. To do this, we used a systematic methodology summarized in Figure 2. We explain the process for collecting and preparing datasets of GPTs and their associated API specifications, the development of a comprehensive taxonomy and the method for detecting personal parameters within API calls, and the analysis framework used to assess the traceability between detected data transmission and stated privacy policies. While our method targets the GPT Store, it can be applied to any LLM app platform where Action-style parameter specifications are available.

3.1 Dataset Collection and Preparation

3.1.1 Initial Data Sources. The OpenAI GPT Store presents only a subset of GPTs, organized into 10 categories. It does not provide a *comprehensive index* of all available GPTs, making exhaustive data collection through this channel alone infeasible. To tackle this limitation, we used a double strategy. We used third-party, community-curated indexes of GPTs, and GPTs datasets collected by other researchers. In particular, we used:

1) GPTStore.ai: From April 5 to April 7, 2024, we scraped GPTStore.ai [66], a third-party website that aggregates listings of GPTs. This crawl returned 96,521 links. We subsequently used these links from April 10 to 12, 2024, to query OpenAI's official store, retrieving metadata for 85,478 GPTs that were publicly accessible. 87% of GPTs

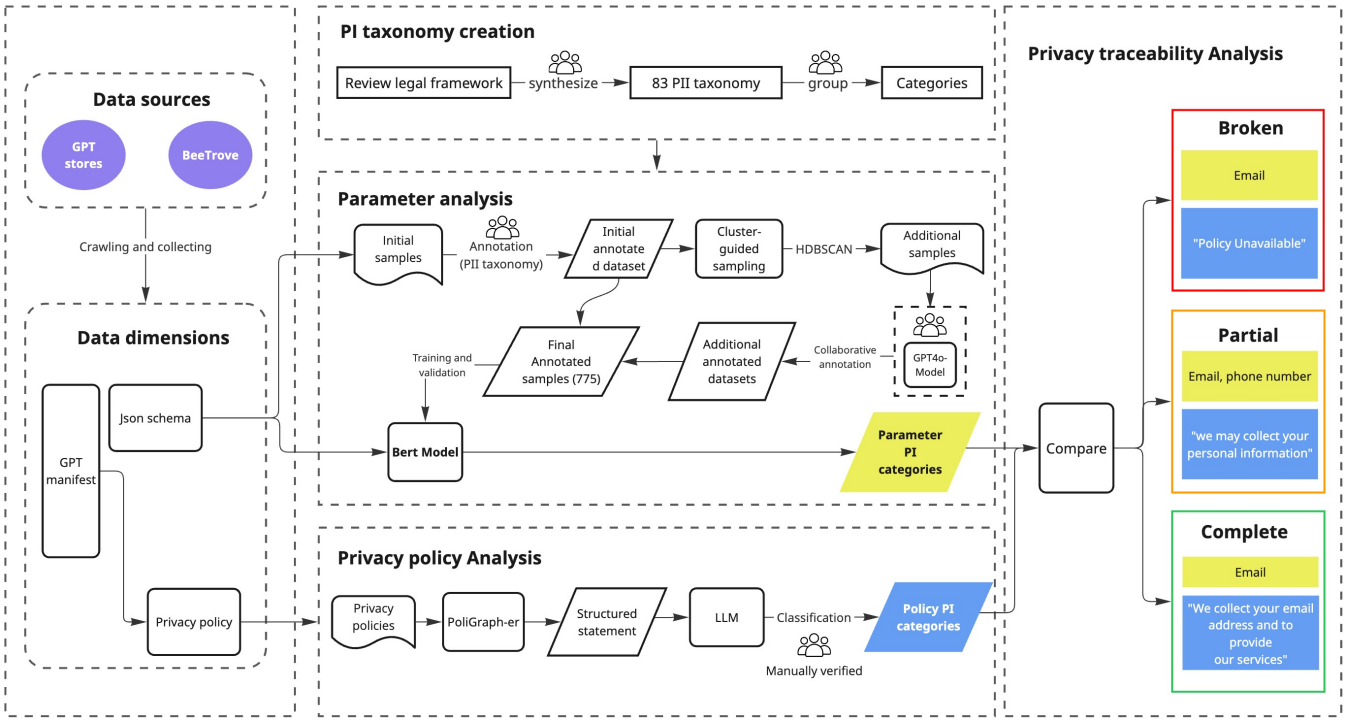

Figure 2: An overview of the method: data collection and analysis

in our dataset remain public, underscoring the timeliness of our findings despite recent API changes.

2) BeeTrove Dataset: to enhance domain diversity and ensure a broader representation across sectors such as productivity, education, finance, and healthcare, we incorporated the BeeTrove dataset [4]. It contains 334,348 entries following a consistent schema.

Deduplicating the overlap between the two sources, the resulting combined dataset comprises 343,122 total unique GPTs. While it does not capture the entire population, estimated to exceed 3 million GPTs [47], it offers a representative sample that reflects the diversity (as we will see later, e.g., in terms of categories).

3.1.2 Collected Data. GPTs are defined by a manifest. We processed each manifest and several key data dimensions within it:

a) Manifest: The manifest is the descriptor associated with the GPT. It contains basic information about the GPTs such as the application id, name, description, query examples and information about the creator. It also defines which tools the GPTs has access to, both provided by OpenAI (e.g., access to files uploaded by the creator, code interpretation, image generation) or external API calls, called actions. Actions include an extended description in the manifest to cover the integration (JSON schema) and a link to the privacy policy. An example manifest is included in Appendix A.

b) Action JSON schema (OpenAPI Specifications): The GPT uses the JSON schema to connect to the external API, and it includes the expected structure of request and response parameters: parameter names, data types, and descriptions. An example action description with a JSON schema and privacy policy link is in Appendix B.

c) Privacy Policy: the privacy policy of each action is obtained through a link included in the action description. The policy should address the data protection and user privacy of the data sent through the corresponding action.

3.1.3 Data Filtering. From our collection of 343,122 unique GPTs, we filtered the dataset following these steps. First, we focused exclusively on GPTs that explicitly defined at least one "action," indicating an integration with an external API, which is the focus of our work. This left 8,192 GPTs that included at least one external API action. Second, we extracted API parameter information from the action JSON schemas provided in their manifests. This step was essential for our classification of potentially transmitted personally identifiable information (PI). Due to variability and occasional inconsistencies in the schema structures, we retained only those GPTs for which parameter data could be extracted in a robust and consistent manner. As a result, our dataset included a final total of 5,286 GPTs that explicitly defined at least one action and from which parameter data could be consistently extracted from their JSON schemas.

3.2 PI Taxonomy Creation

To detect specific types of personal data being collected through API parameters, we need a data taxonomy for personally identifiable information (PI). We first considered different legal frameworks to guide our efforts. For example, in the United States, the *National Institute of Standards and Technology* (NIST) provides official guidance on the classification and handling of personally identifiable information [36], while in the European Union, the *General Data*

Protection Regulation (GDPR) mandates strict controls on data minimization, lawful data processing, and user rights [17]. Although these frameworks differ in their specific details and jurisdictions, they share foundational principles: transparency, individual consent, and effective data security. However, these regulations do not provide a detailed definition of PI categories or a data taxonomy, as it has been pointed out on previous research [40].

To overcome this limitation, we adapted a taxonomy of PI [58] inspired by the work of [40, 61]. Since the taxonomy is primarly inspired by US regulation, we extended their work by including the data specifications outlined in the GDPR. More specifically, we incorporated both the general data definitions from Article 4 as well as the regulations on the processing of special categories of personal data from Article 9. This includes *Online Identifiers for Profiling and Identification*, *Genetic Data*, *Health Data*, and *Sensitive Personal Data* such as race, religion, sexual preference, or law enforcement data.

We also adapted the taxonomy to fit the specifics of conversational AI. For example, many GPTs collect the whole message from the user or conversation, instead of individual parameters. We defined the category of CONVERSATION to group such cases. In our analysis pipeline, parameters falling into this category trigger an aditional scrutiny to assess what type of data that specific GPTs may be collecting.

As a result, we use an 83-categories PI taxonomy that can be used for privacy analysis in the context of LLM app integration. The complete data taxonomy is included in Appendix D.1. Our classification scheme improves established data-privacy literature with its granularity. For instance, many prior works lump personal identifiers together or focus solely on classic fields like name, address, or date of birth. By contrast, our extended taxonomy captures both newly emerging identifiers (e.g., CRYPTO wallet addresses, ONLINE_IDENTIFIERS, WORK_PHONE_NUMBER) and highly sensitive medical elements as defined in the GDPR. This granularity is crucial for ensuring that LLM apps handle data responsibly across a wide array of personal information types.

3.3 PI Classification in API Parameters

We developed a systematic classification approach to identify Personal Information (PI) transmitted via third-party APIs in GPTs. This involved preparing the API parameter data for classification, annotating a subset of this data according to our PI taxonomy, and then training and validating a machine learning model to automate this classification at scale.

3.3.1 Preparation of Input Data for Classification. For each API parameter identified in our dataset of GPTs, we extracted relevant textual information to serve as input features for our PI classifier. As detailed in § 3.1.2b, this included the parameter name (e.g., "email"), the parameter description provided by the creator (e.g., "Email address of the user making a booking"), and the contextual information from the GPT name (e.g., "Virtualdeborah") and GPT description (e.g., "Book complex tasks from a real human"). These distinct textual fields were concatenated into a single input string to preserve the contextual richness of each parameter, enhancing the quality of subsequent annotation and classification.

3.3.2 Annotation of API Parameters. With the input data prepared, we proceeded to annotate a subset of the API parameters using the 83-category PI taxonomy developed in § 3.2. This annotation process was multi-staged to ensure quality and address data imbalances. We began by randomly selecting 300 parameters from distinct GPTs to form an initial annotation set. These parameters were independently annotated by three of the paper authors, all previously trained in privacy-related data classification, using our PI taxonomy. Across the annotated parameters, a total of 24 distinct PI categories were observed. The first 100 parameters yielded a Krippendorff's Alpha of 0.5897, indicating moderate agreement and highlighting areas for improvement. Following detailed discussions and refinements to the annotation guidelines to address ambiguities. For instance, initial disagreements arose from: (1) classifying location-related parameters with varying granularity (e.g., *COUNTRY* vs. *GEO_LOCATION*); (2) interpreting generic parameter names (e.g., user_input) that could capture user conversation; and (3) distinguishing between closely related PI types (e.g., *SCREEN_NAME* vs. *ONLINE_IDENTIFIERS*). After this, the remaining 200 parameters were annotated, resulting in an improved Krippendorff's Alpha of 0.8379. This demonstrated a high level of agreement and established consistent criteria for identifying PI within API parameters.

3.3.3 Enriching the Dataset to Address Class Imbalance. Analysis of the initial annotation set revealed considerable class imbalance across the 83-category PI taxonomy, with many categories sparsely represented or entirely missing. To address this limitation, we applied two complementary dataset enrichment strategies:
1) Cluster-guided Sampling via Broader Category Groupings: We sought broader coverage of related PI labels by grouping the 83-category taxonomy into higher-level categories (e.g., Basic Personal Information, Contact Information, Government or Official IDs; see Appendix D.2). A preliminary model, trained on the initial 300 annotated samples, generated predictions on the remaining unlabeled data. Using HDBSCAN [39], we then clustered the unlabeled parameters within each major category. From each cluster, we selected the most representative examples, thereby boosting our coverage of diverse PI samples while limiting redundancy. This process yielded 213 additional samples for manual annotation, distributing 20 per category except for the Government or Official IDs category, which had only 13 suitable samples.
2) Targeted Expansion through Human–GPT-4o Collaboration: To further enrich low-frequency classes, we employed GPT-4o in a collaborative annotation workflow inspired by frameworks such as MEGAnno+ [28] and CoAnnotating [32] among others [68]. Using the initial 300 annotations as a base, we applied TextGrad [81] to optimize a task-specific prompt for GPT-4o, enhancing its ability to classify PI categories. We ran GPT-4o across the corpus and selected parameters that it identified as PI. To strategically surface cases likely to improve coverage of rare categories, we compared GPT-4o's findings against those of an interim classifier trained on the 513 manually labeled samples already gathered (from the initial 300 and the 213 from cluster-guided sampling). Parameters identified by GPT-4o but missed by this interim classifier were prioritized. From each newly identified class through this process, we selected up to 20 samples for manual annotation, yielding 262 additional labeled parameters.

Together, these strategies produced a final **annotated dataset**[1] **for training our classification model**, with 775 API parameters spanning 54 unique PI categories — which are the 54 first shown in Table 7 in the Appendix. This enriched dataset significantly improved class balance and served as a robust foundation for training our final classification model. The Krippendorff's Alpha for this expanded 775 sample dataset was 0.8147.

3.3.4 Qualitative Analysis. In addition, during the annotation process, we took memo notes on noteworthy cases involving sensitive personal information or problematic behavior. These cases were discussed among researchers. We qualitatively reviewed and reflected on these examples to provide additional context for the types of data being collected and to reflect on their appropriateness and potential privacy risks. This is the basis for the qualitative insights shown in § 4.1.3.

3.3.5 Training and Validation. We use Fastfit to fine-tune a classifier based on Roberta-Large [34], a pre-trained BERT model. FastFit [79], a method specifically designed to deliver fast and accurate few-shot classification, particularly in scenarios with many semantically similar classes, underscoring its suitability for the complex classification challenges present in our study. To evaluate classifier performance, we employed a stratified 5-fold cross-validation strategy to preserve class distributions across training and validation splits. We obtain a macro F1 score of 87%. Detailed analysis revealed consistently high precision and recall for most categories, including near perfect performance (precision and recall >92%) for critical categories such as *MEDICAL_HISTORY*, *EMAIL_ADDRESS*, and *CV_RESUME*. See Table 13 in the Appendix. The validation results confirm the effectiveness and reliability of the classifier, demonstrating robust performance across different contexts and data categories. Despite small classification errors between closely related categories (which would still be useful to know *there is* personal information involved in that parameter), the high scores across multiple evaluation metrics confirm the robustness and scalability of our methodology.

Alternative methods for classification: We also considered using an LLM (GPT-4o) for this classification task. However, as we show in Table G.1 (for GPT-4o) and Table G.2 (for our fine-tuned RoBERTa model), our method based on the fine-tuned BERT architecture (RoBERTa) worked much better for the actual PI classification. Specifically, while GPT-4o achieved a Macro F1-score of 0.61, our fine-tuned RoBERTa model demonstrated significantly higher performance with a Macro F1-score of 0.87. This finding is consistent with existing literature, which suggests that it is often better to use smaller language models fine-tuned on specific tasks rather than relying on zero-shot capabilities of larger LLMs for classification [8, 16]. Moreover, smaller, fine-tuned models can offer significant resource efficiency advantages while maintaining strong performance [60]. Thus, while GPT-4o was actually very useful to expand our dataset (see § 3.3.2) for a more balanced training set, our fine-tuned RoBERTa-based approach provided superior accuracy and reliability. In addition, by leveraging RoBERTa-large, a fine-tuned transformer model with 355 million parameters requiring approximately 710 MB in FP16 precision, we achieve strong classification performance without the need for massive infrastructure. In contrast, larger foundation models such as GPT-4o (200 billion parameters, 400 GB) or GPT-4o-mini (16 GB) impose considerable memory and computing overhead.

[1]The annotated dataset will be provided on acceptance of the paper

3.4 Traceability Analysis

Our traceability analysis framework evaluates the alignment between the personal data transmitted by GPTs to third-party APIs and the disclosures made in their associated privacy policies. This process involves enhancing the capabilities of PoliGraph-er [13], an automated policy analysis tool, with Large Language Models (LLMs) to bridge the gap between policy statements and our specific data classification taxonomy.

3.4.1 PoliGraph-er Processing of Privacy Policies. To examine how creators/organizations disclose the types of personal information (PI) they collect, we analyzed the text of privacy policies using PoliGraph-er [13], a state-of-the-art NLP tool designed to extract structured information about data practices from privacy policies. It is worth noting that not all GPTs had valid privacy policy links, as we detail in the results with specific figures: some links are unusable, point to unrelated content or cannot be processed by PoliGraph-er — e.g., non-English privacy policy, malformed HTML etc. For all the valid privacy policies, PoliGraph-er can extract structured statements about data collection, enabling us to analyze which PI types were explicitly disclosed. These graph-based outputs form the foundation for comparing policy statements with actual data transmitted in API interactions, as detailed next.

3.4.2 Extending PoliGraph-er with LLMs for Semantic Clustering. While PoliGraph-er effectively extracts structured statements about data collection from privacy policies (e.g., "We collect → email"), directly comparing these outputs with our PI classification of API parameters (see § 3.3) presents a challenge. The extracted data type mentions (such as "email address," "location data," or "user's job title") are expressed in diverse natural language and do not directly align with the categories in our PI taxonomy. To enable meaningful comparisons between what policies claim to collect and what is actually transmitted through APIs, we developed a process to semantically cluster these extracted data type mentions and map them to our standardized PI taxonomy. This crucial step translates the diverse language of privacy policies into a consistent, structured format, facilitating direct traceability analysis. To implement this, we first extracted 26,629 "collection" relationships from successfully parsed policies, each containing an entity, a collection relationship, and a data type. After deduplication, this yielded 5,741 unique data type mentions. We then employed GPT-4o-mini to perform the semantic clustering and mapping. For example, policy statements referencing *"position"*, *"professional title"*, or *"expertise"* were grouped by the LLM and mapped to our *JOB_TITLE* category.

To evaluate the performance of this LLM-assisted mapping, we sampled up to 10 mapped statements per PI category, producing a validation set of 264 examples. A researcher manually verified whether each LLM classification matched the correct taxonomy label. GPT-4o-mini achieved an overall accuracy of 87.83% in this

task, demonstrating strong alignment with human judgment in clustering and categorizing these collection statements — the specific prompt used for this categorization is provided in Appendix E.2 , with the corresponding confusion matrix shown in Table 10.

3.4.3 Compliance Types. To assess privacy compliance in GPT third-party integrations, we follow the literature on traceability in other domains, such as social media, smartphone apps, and voice assistant skills [3, 14, 42, 80, 86], adapted to suit the data flows characteristic of this ecosystem. Each GPT is assessed as having *broken*, *partial*, or *complete* traceability [3, 42, 80, 86]. These classifications are defined as follows:

Complete: occurs when all types of personal data transmitted by the GPT, according to its OpenAPI schema or parameter definitions, are explicitly mentioned in the privacy policy. For instance, if a GPT collects a user's email address and location, a statement such as *"We collect your email address and geographic location to provide our services"* is considered fully traceable.

Partial: applies when only some of the data types transmitted via parameters are mentioned in the policy, or when the references to data collection are **vague** or generic. For example, if a GPT transmits both email address and phone number, but the privacy policy only includes the phrase *"we may collect your personal information"*, this would constitute partial traceability. Similarly, policies that mention one of the collected data points (e.g., email) but omit others (e.g., phone number or location) fall into this category. Some GPTs rely on a parent organization's privacy policy instead of offering GPT-specific data handling details. We classify these as partial disclosures: they reference data practices in general terms but lack the direct, GPT-specific description of data flows needed for a complete disclosure.

Broken: refers to cases where the GPT lacks a privacy policy, provides a broken or inaccessible link, or includes a policy that does not mention any relevant data types, despite transmitting them. For instance, a GPT that collects geolocation to check for the weather APIs but its policy does not mention location.

3.5 Active Audit: Data Sent to Third Parties

To complement our static analysis of Action schemas, we conducted an active audit to directly observe the data transmitted by GPT-integrated APIs to third-party endpoints. We implemented three minimal services, each corresponding to an Action from our dataset, and deployed them. Each service simply printed to the console the parameters received in the request, allowing us to verify the payload content. The three Actions reproduced common cases of potential exposure: (*i*) **Weather-Echo**, a weather query Action with a single *location_data* parameter (Fig. 7); (*ii*) **Conversation-Echo**, an Action with a *original_text* parameter of type CONVERSATION (Fig. 6); and (*iii*) **LeetCV**, a multi-field Action requesting contact details, salary, past_employments, job title, email, CV... (Fig. 5). For each case, we interacted manually with the corresponding GPT, providing controlled inputs and observing the exact parameters received by our servers. This setup allowed us to confirm whether parameters were populated beyond the user-provided information, whether the data sent matched the content of the prompt, and whether user conversation could be extracted. This active audit serves to validate and complement our large-scale analysis by providing empirical evidence of the actual data transmitted at the transport layer.

4 Results

4.1 Personal Data Parameters

Building upon the procedure described in § 3.1.3, we extracted a total of 44,102 API parameters. These parameters, defined in creator-supplied action schemas, offer insight into how user input may be transmitted to external services. **Our analysis across 44,102 API parameters** reveals that 15,392 entries, approximately **35% correspond to categories of sensitive or identifying data** (See Table 1). While a majority of parameters fall into the non-sensitive class, a considerable portion involve fields with privacy implications. Parameters classified as CONVERSATION, representing ambiguous or obfuscated fields that may extract entire prompts or conversation history, **account for 11.5% of all cases**. Web addresses (URL) and language metadata (LANGUAGE) are each found in about 5.8% and 5.7% of parameters respectively, sometimes linking to external user content. Email addresses appear in 1.5% of parameters, while passwords are detected in 1.8%. Information related to geographic location is present in roughly 0.9%, and professional roles such as job titles and résumé content appear in 0.35% and 0.3% of cases respectively. Other noteworthy data types include phone numbers and medical history, each found in approximately 0.2% of parameters. Although individually these figures may seem modest, they represent thousands of individual instances where personal data may be exposed through GPT integrations.

Certain APIs further highlight the privacy risks of unregulated integration. For example, the GPT named *LeetCV - Online Resume Builder* [31] exposes endpoints that request full names, email addresses, and authentication credentials. Similarly, we found a GPT integrated with an API for email automation, enabling the transmission of recipient identities and unstructured message content that may contain sensitive information. These integrations demonstrate how even commonplace development choices can result in the inadvertent collection of high-risk user data, especially when input parameters are not properly sanitized or when API documentation lacks transparency.

4.1.1 Personal Parameter across Categories. A closer look at the distribution of these parameters across application categories shown in Figure 3 reveals heterogeneity in the privacy risks posed by different GPT categories. Notably, the *lifestyle*, *productivity*, and *other* application categories collectively account for the largest concentration of PI parameters. Within these categories, sensitive fields such as *EMAIL_ADDRESS*, *LANGUAGE*, *JOB_TITLE*, and chat extraction parameters (*CONVERSATION*) emerge frequently, raising further privacy concerns.

4.1.2 Distribution of Personal Data Parameters in GPTs. Figure 4 illustrates the number of distinct PI parameters (i.e., parameters not classified as *NON_PI*) used by GPTs. The distribution is top-heavy: the annotation on the plot shows that fully 90% of GPTs specify no more than two PI parameters. The dashed green and red vertical lines mark the median (2.0) and mean (2.2), underscoring how tightly usage is clustered at the low end. In absolute terms,

Table 1: Category counts over 44,102 API parameters

Category	Count
NON_PI	28,710
CONVERSATION	5,083
URL	2,568
LANGUAGE	2,493
PASSWORD	779
EMAIL_ADDRESS	654
COUNTRY	600
DATE_TIME	524
GEO_LOCATION	380
ONLINE_IDENTIFIERS	362
ZIPCODE	264
SCREEN_NAME	231
PERSON	172
JOB_TITLE	152
BIRTH_DATE	148
CV_RESUME	134
CRYPTO	131
SHOPPING_BEHAVIOR	112
ADDRESS	98
MEDICAL_HISTORY	98
PHONE_NUMBER	93
ACTIVITIES	63
PLACE_OF_BIRTH	47
PERSON_GENDER	43
EDUCATION_INFORMATION	36
VEHICLE_REGISTRATION_NUMBER	32
HOME_ADDRESS	32
PICTURE_FACE	26
NATIONALITY_CITIZENSHIP	13
PERSON_HEIGHT	9
PERSON_WEIGHT	9
NUMBER_OF_CHILDREN	5
HEALTH_INSURANCE_ID	1

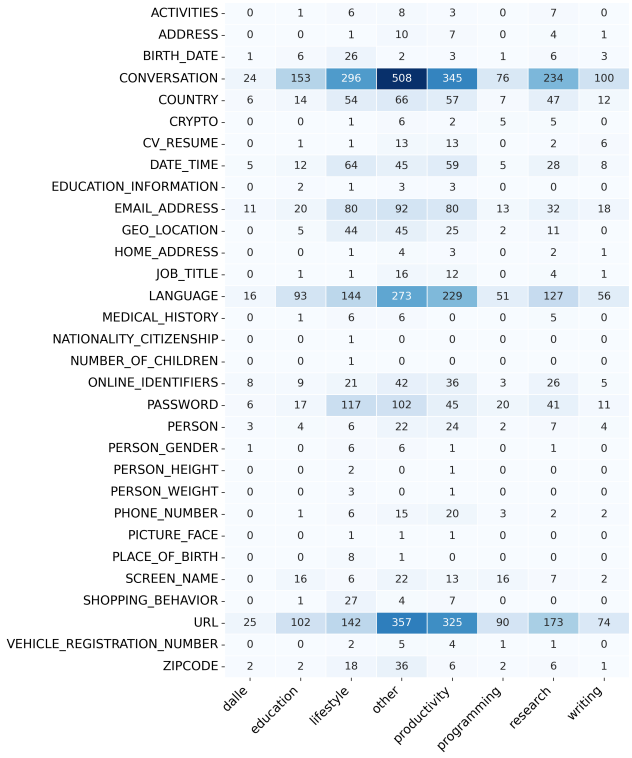

Figure 3: Heatmap of the frequency of PI categories in API parameters (columns) by GPT category (rows)

1,722 GPTs rely on a single PI parameter, while 601 use exactly two. Interestingly, the count rises again to 1,317 for three parameters. From that point the numbers decline sharply, but they still show a very sizable number of GPTs having more than four parameters: 79 GPTs specify four parameters, 32 specify five, and 128 specify six. Altogether, 3,879 GPTs, about 98% of the sample, use six or fewer parameters. Beyond six, the tail is very thin: 47 GPTs use seven parameters, six use eight, and so on.

4.1.3 Unpacking Creator Practices in API Parameter Definitions. Our quantitative analysis revealed extensive data collection across a broad spectrum of GPTs. To gain a deeper understanding of the specific types of data being collected and assess whether they are gathered for appropriate purposes, this section presents illustrative cases that exemplify excessive data collection practices in GPTs.

Creator misconceptions about data handling: Some creators appeared to assume that sensitive user data would be redacted before reaching their APIs, an expectation reflected in their parameter documentation. For example, one creator defined a parameter as *"Parameter Name: original_text"*, indicating that they were collecting the conversation. In the parameter description, they specified: *"Please provide the original request (only containing user input) that triggered the API call, as this information will be used to improve the performance of the API. If the text contains sensitive user data, such as names, please redact them as **."*. Such instructions suggest creators' assumption that either the GPT model or the platform would handle the redaction of sensitive information before it reaches the API, without considering themselves responsible for safeguarding user

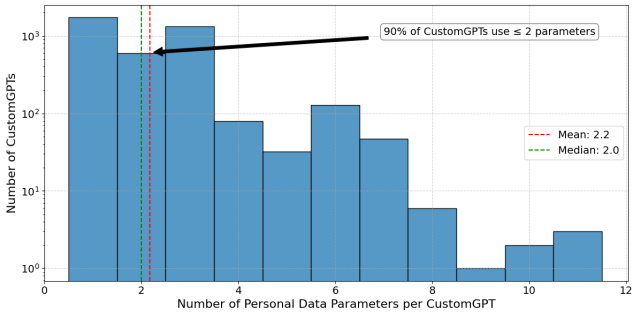

Figure 4: Histogram showing the distribution of personal data parameters used across GPTs

data. In practice, however, GPT does not perform this preprocessing, and full conversations are still passed to creators (see § 4.5).

Sensitive information collection: Some GPTs are designed to support task contexts that involve sensitive information, and many of them collect personal data through their API parameters. For example, in a CV-generation GPT [31], the parameters included fields for the candidate's email address and phone number. Payment-related GPTs also highlight similar concerns. These applications may assist users in managing their personal financial status and request highly sensitive information such as bank account numbers and invoice

details. In the health domain, GPTs offering health insurance suggestions frequently include parameters for country-specific insurance numbers or health system records and identifiers. Moreover, some GPTs collect information not only about the user themselves but also about other users. For instance, one GPT designed to scrape social media history requested details such as names and phone numbers of targeted individuals and could return complete social media histories.

Excessive data collection: Beyond the sensitivity of individual data fields, some GPTs collect excessive amounts of personal information across multiple categories. As shown in Figure 4, several GPTs request not only highly sensitive data but also a wide range of it within a single app. For example, one GPT designed to assist with hiring or job searching simultaneously requests ten types of personal information, including a user's full name, email address, phone number, location, and salary. In another case, a medical assistant GPT collects eleven types of personal information through API parameters, including patient ID, hospital location, birthdate, home address, phone number, email address, and full name.

4.2 Co-occurrence and Cross-Contextual Risks

While previous sections focused on the privacy implications of individual GPTs and their data handling practices, we now turn to the systemic risks that may arise from shared Actions and co-occurrence patterns across GPTs. This ecosystem-level view helps surface potential for indirect data exposure and cross-contextual user tracking. While most GPTs (89.8%) rely on a single Action, a non-negligible portion (10.2%) incorporate two or more, revealing more complex integrations. Among these multi-Action GPTs, 72.3% connect to endpoints in entirely different domains, an architectural choice that increases the likelihood of indirect data exposure through cross-contextual linkages. Co-occurrence analysis further supports this concern: nearly half of all Actions (47.3%) are used in GPTs that invoke at least one other Action, suggesting an ecosystem where data from distinct sources may be collated or inferred. Prominent Actions like WebPilot, Gapier, and AdIntelli exhibit especially high weighted degrees of co-occurrence, often appearing together in GPTs that span categories such as productivity, research, and media. For instance, WebPilot co-occurs with over 60 other Actions, including Zapier, AdIntelli, and GitHub APIs, hinting at latent pathways for user data to traverse across otherwise disconnected services. These overlapping integrations present a systemic risk: even if individual Actions appear compliant in isolation, their joint usage can enable composite user profiling or inadvertent leakage of sensitive data across domains.

4.3 Usage of GPTs with PI Parameters

In Table 2, we aggregate usage (total number of whole user conversations with GPTs) across all GPTs that request a given PI parameter. For example, GPTs collectively sending CV_RESUME through Actions surpass 900,000 conversations with users in total, clearly demonstrating that may affect a substantial audience. Similarly, GPTs requiring *EMAIL_ADDRESS*, *PERSON*, or *PHONE_NUMBER* each accumulate hundreds of thousands of user conversations once all such GPTs are combined. In contrast, while GPTs with *MEDICAL_HISTORY* parameters see relatively fewer overall conversations, their sensitive nature remains noteworthy and is still in the thousands. Thus, these sensitive categories still reach considerable numbers of interactions with users, reinforcing the importance of privacy safeguards for both highly used and specialized GPTs.

Table 2: GPT conversations by category

Category	GPT Conversations
CV_RESUME	972,649
EMAIL_ADDRESS	346,540
PERSON	219,343
PHONE_NUMBER	170,853
EDUCATION_INFORMATION	56,504
MEDICAL_HISTORY	8,258

4.4 Traceability Results

4.4.1 General Traceability Results. Out of the 5,286 GPTs analyzed, 792 GPTs (≈ 15%) exhibited **complete** traceability, 2,258 GPTs (≈ 42.7%) exhibited **partial** traceability, and 2,004 (≈ 37.9%) exhibited **broken** traceability. PoliGraph-er could not process 232 GPTs because their privacy policies were not in English, and they were therefore excluded from the traceability analysis. Digging deeper, there were 3,400 GPTs for which the policy content was made available for comparison against their the PI collected through the APIs they connect to. Many of these policies, however, were privacy policies that had been reused across several GPTs. After de-duplication, there was a total of 2,140 unique privacy policies. We provide further analysis on them in the next sections.

4.4.2 Automated Policy Generators. Several broader trends emerged from the analysis of the privacy policies. Many GPT creators appeared to rely on automated policy generators, particularly from platforms such as *privacypolicyonline.com* (379 instances), *freeprivacypolicy.com* (180), *iubenda.com* (14), and *termly.io* (9). We determined this based on the host domains of the privacy policies, which directly matched the domains of the policy generator service. These often produced generalized and sometimes inaccurate disclosures. Some creators hosted privacy policies via plain HTML documents or by linking directly to official legislative texts (e.g., *canada.ca* with 142 occurrences). Moreover, the use of generic, third-party policies was common: *google.com* (388), *github.com* (155), and *azurewebsites.net* (51) were cited, which typically were not tailored to a GPT's unique data practices.

4.4.3 API Integrators. A big proportion of GPTs rely on API integrators such as *zapier.com* (2,409) and *gapier.com* (3,874, a Zapier-related domain), which act as middleware platforms that facilitate connections between thousands of third-party services. These platforms often allow a single API call to trigger workflows that dynamically invoke multiple downstream APIs, complicating the traceability of data flows. However, the linked privacy policies are typically broad and generic, offering limited insight into what data is actually collected, shared, or stored during execution. This lack of specificity poses a significant privacy risk, as users may be unaware of the full extent of the data processing involved. Other aggregators included *rapidapi.com* (313) and *pluginport.io* (219), while domain-specific APIs such as *api-football.com* (118) and *weather.gov* (113)

Table 3: Compliance breakdown by parameter requested

Category	Broken	Complete	Partial
ACTIVITIES	14	0	19
ADDRESS	18	0	16
BIRTH_DATE	44	1	11
CONVERSATION	750	10	1677
COUNTRY	117	0	197
CRYPTO	30	0	6
CV_RESUME	28	0	16
DATE_TIME	110	0	166
EDUCATION_INFORMATION	9	0	3
EMAIL_ADDRESS	231	13	154
GEO_LOCATION	76	0	105
HEALTH_INSURANCE_ID	0	0	1
HOME_ADDRESS	10	0	7
JOB_TITLE	27	0	16
LANGUAGE	212	0	1222
MEDICAL_HISTORY	11	0	10
NATIONALITY_CITIZENSHIP	4	0	0
NON_PII	1662	753	2103
NUMBER_OF_CHILDREN	0	0	2
ONLINE_IDENTIFIERS	120	3	69
PASSWORD	288	2	183
PERSON	61	3	36
PERSON_GENDER	16	0	7
PERSON_HEIGHT	6	0	1
PERSON_WEIGHT	5	0	2
PHONE_NUMBER	41	1	20
PICTURE_FACE	5	0	2
PLACE_OF_BIRTH	6	1	4
SCREEN_NAME	68	0	39
SHOPPING_BEHAVIOR	22	0	28
URL	487	8	1318
VEHICLE_REGISTRATION_NUMBER	9	0	6
ZIPCODE	53	0	41

were also observed. Additionally, *webpilot.ai* appeared in 8,075 instances, functioning as an intermediary for injecting external web content into GPT interactions.

4.4.4 Traceability by Requested Parameter. Table 3 provides a detailed breakdown of broken, complete, and partial disclosures by the specific parameter requested (e.g., phone numbers, precise geolocations). A recurring concern was that many API parameters transmitted personal data types not mentioned in their associated privacy policies. Failure to reference frequently requested parameters (e.g., phone numbers, emails, geolocation) places end-users at potential privacy risk. These omissions highlight the urgency of stronger audit mechanisms and more transparent design practices.

4.4.5 Association Between Data Categories and Privacy Policy Completeness. We investigated whether the specific types of information handled by GPTs are associated with differences in privacy-policy completeness, as measured by the compliance label (*complete*, *partial*, or *broken*). To assess statistical relationships between data types and policy completeness, we ran Chi-squared tests of independence [38] and then calculated Cramér's V to quantify the strength of association [12]. Effect-size interpretations follow Cohen's criteria for Cramér's V with two degrees of freedom [11]: negligible ($V < .07$), small ($.07 \leq V < .21$), medium ($.21 \leq V < .35$), and large ($V \geq .35$). Unless noted otherwise, all reported associations are statistically significant at $p < .05$.

A large association was observed for *Non-Personal Information*. Of the 792 GPTs with complete traceability, a substantial majority (753) are GPTs where no personal information was found to be sent via the API parameters (i.e., all parameters for that action were classified as *NON_PI*). In addition, six categories exhibited smaller associations: *Demographic or Social Attributes*, *Information Extracted from the Chat* (CONVERSATION category), *Online and Digital Identifiers*, *Employment and Professional Information*, *Location and Address Information*, and *Contact Information*. These data types often require nuanced handling and may introduce uncertainty or hesitancy in disclosure, resulting in less comprehensive policies. All remaining categories showed negligible effect sizes.

4.4.6 Post-hoc Pairwise Comparisons. To pinpoint where the global χ^2 effects originate, we followed Sharpe's "compare–cells" methodology [63]. A "cell" refers to the count of GPTs for a given combination of PI category and it's traceability assesment. Specifically, we ran pairwise two-proportion z-tests comparing the presence vs. absence of each PI family category (as defined in Appendix D.2) across all three compliance levels. Tests were conducted only where all expected cell counts exceeded 5. The Bonferroni method was applied to adjust p-values for multiple comparisons ($\alpha = 0.05$). Table 4 summarizes the PI families with an overall significant association (from the Chi-squared test) and lists the specific pairwise contrasts that remained significant after this correction, with Δ indicating absolute percentage-point differences and *OR* representing odds ratios.

Table 4: Cramér's V association & significant post-hoc z-tests

Family	$p < 0.05$	Cramér's V	Compliance Level	Δ (pp)	OR
Basic Personal Information	✓	0.0283	broken ↑	+8.3	2.09
			partial ↓	−5.6	0.71
			complete ↓	−2.7	0.66
Non-Personal Information	✓	0.3992	broken ↓	−1.0	0.88
			partial ↓	−22.5	0.23
Demographic / Social Attr.	✓	0.1593	partial ↑	+17.5	8.69
Information from Chat	✓	0.1589	broken ↓	−1.9	0.79
			partial ↑	+12.6	3.02
Online & Digital IDs	✓	0.1437	complete ↓	−9.9	0.06
			broken ↑	+3.0	1.38
			partial ↑	+6.9	1.69
Employment / Professional	✓	0.0808	broken ↑	+30.4	6.48
			partial ↓	−21.8	0.33
Location / Address	✓	0.0760	partial ↑	+3.4	1.29
			broken ↑	+5.6	1.72
Time Information	✓	0.0443	partial ↑	+7.4	1.84
			broken ↑	+1.4	1.17
Contact Information	✓	0.0699	complete ↓	−2.8	0.65
			broken ↑	+12.6	2.78
			partial ↓	−9.8	0.57
Financial / Payment	✓	0.0509	partial ↓	−32.1	0.22
			broken ↑	+40.7	9.73
Education Information	✓	0.0360			
Behavioral, Activity and Web Tracking	×	0.0238			
Health and Medical Information	×	0.0203			
Government or Official IDs	×	0.0161			

Severely incomplete disclosures tend to cluster around financial and employment data. GPTs that process Financial and Payment Information or Employment and Professional Information are dramatically more likely to feature broken policies, with odds ratios

Table 5: Compliance breakdown by GPT category

GPT Category	Broken	Complete	Partial
Other	407	193	510
dalle	26	17	25
education	101	29	138
lifestyle	290	82	218
productivity	256	110	336
programming	95	40	68
research	210	106	213
writing	65	23	87

of 9.73 and 6.48 respectively. These same categories also show significant drops in the rate of even partial compliance, highlighting their consistent association with policy failures.

Basic identity traits also appear to reduce overall policy quality. The inclusion of Basic Personal Information correlates with a decrease in complete policies by 2.7 percentage points and an increase in broken ones by 8.3 points. This suggests a broad reluctance or inability to comprehensively disclose practices related to core identity attributes.

Conversational and demographic content appears to prompt superficial compliance. GPTs incorporating Information extracted from the chat or Demographic and Social Attributes are considerably more likely to have partial policies, with odds ratios of 3.02 and 8.69, respectively. These patterns suggest that while creators may acknowledge the need for some form of disclosure, they often stop short of full transparency.

In contrast, GPTs that only process Non-Personal Information are associated with markedly better outcomes. The likelihood of partial compliance decreases by 22.5 percentage points in this group, and broken policies become slightly less common as well. Finally, temporal and geospatial signals introduce subtler effects. GPTs that handle Time Information or Location and Address Information show modest but consistent increases in both broken and partial disclosures. These results may reflect underlying ambiguity or creator discomfort around documenting practices related to tracking or situational data.

4.4.7 Traceability by GPT category. Table 5 shows that traceability issues permeate many GPT usage categories. While the "Other" category exhibited the largest share of broken disclosures (407), "research" and "productivity" services also showed high numbers of partial or broken policies. These findings illustrate that policy incompleteness is not confined to any single application but is widespread across diverse GPT use cases.

4.4.8 The Good, the Bad and the Ugly creators. In total, there are 2,175 unique creators of GPTs requesting parameters in their calls to an external API:

The good: There are 443 creators whose GPTs are all fully traceable, representing 17% of all creators. These creators clearly articulate and justify the permissions they request in their privacy policies. For example, creator Sora AI has 18 complete GPTs, and creator Crypto maintains 13 complete GPs, both exemplifying best practices in disclosing PI collection through API parameters.

The bad: There are 957 creators with all GPTs with broken traceability, representing approximately 44% of all creators. These creators provide inadequate or no explanation of permissions in their privacy policies. For instance, aikitcentral.com has 193 partial and 5 broken GPTs, showing widespread inconsistency, while Finntech1 has 34 partial GPTs with no clear justifications provided.

Table 6: Creator compliance distribution

Compliance Type	Creators
Partial only	923
Broken only	957
Complete only	443
Both Partial and Complete	62

The ugly: There are 923 creators with partial traceability, representing 42% of all creators. Their privacy policies partially address the permissions requested, indicating incomplete or unclear disclosure. A representative case is tinycorp.ai, which has 43 broken GPTs, reflecting a high level of undocumented permissions. Similarly, one creator manages 42 GPTs with partial traceability, and BREEBS has 35 GPTs that are inconsistently documented. In addition, there is a small group of 62 creators (around 2%) whose GPTs are mixed, with both partial and complete traceability practices.

Table 6 summarizes these creator groupings. The results show a widespread problem of inadequate compliance practices, with 'bad' and 'ugly' creators significantly outnumbering those with good compliance. This suggests the potential of targeted interventions aimed at improving creator compliance across the board.

4.5 Findings from the Active Audit

Our active audit produced three main observations: **1. Location (Weather-Echo).** The request body contained only the *location_data* field, populated exactly as written in the user prompt. We did not observe automatic inclusion of IP-based geolocation, derived coordinates, or coarse location fallbacks. When no location was provided in the prompt, the parameter was either omitted or set to an empty string. This confirms that location values originate from the prompt rather than from other metadata. **2. Conversation (Conversation-Echo).** When a parameter such as *original_text* was defined in the schema, the request body included the full user prompt without any redaction or anonymization, even when the input contained personal information. This occurred despite the parameter description explicitly instructing the GPT to remove or mask sensitive elements, such as names. In the absence of such a parameter, no conversation text was transmitted. These findings reinforce our classification of CONVERSATION parameters as high-risk. **3. Over-collection (LeetCV).** For schemas listing multiple PI categories, the model attempted to populate all available parameters with information extracted from the prompt, leaving empty values for missing fields. This indicates that over-collection risk arises from the breadth of parameters defined by creators.

5 Discussion

5.1 Main Takeaways

5.1.1 Massive Personal Data Collection with no Transparency. Our analysis of 5,286 GPTs that integrate via APIs to services outside OpenAI's infrastucture, reveals a total of 44,102 API parameters used by those GPTs, out of which approximately 35% correspond to categories of sensitive or identifying personal data that are being

collected. Even more worryingly, when analyzing the traceability with the GPTs' privacy policies, the results are that only 792 GPTs (15%) exhibit complete traceability, i.e., they disclose in their privacy policy the personal data that is collected through the API parameters. This paints a bleak picture of the current state of the GPTs ecosystem when it comes to personal data being taken out of the OpenAI ecosystem and the lack of transparency given to users explaining that.

5.1.2 Validation via active audit. The active audit results reinforce our main conclusion: the primary exposure vector is the combination of what creators request in the JSON schema and what the model extracts from the user prompt, rather than metadata added outside the prompt. Risk increases notably for CONVERSATION parameters and for schemas that combine multiple personal information fields, both of which are common in our corpus. These findings suggest potential mitigations at the platform level, such as editable previews of outbound payloads and schema linting to discourage overly broad parameter definitions. Our tests did not find evidence of session correlators in the examined cases, but we cannot exclude their presence in other Action types or under different authentication contexts.

5.1.3 Privacy Policy Completeness & Parameter Sensitivity. Our findings reveal that while privacy policy completeness is not uniformly tied to data sensitivity, GPTs that process conversational content or user-identity attributes are associated with notably less complete disclosures. Surprisingly, highly sensitive categories (e.g., health or government ID information) did not show strong associations, possibly due to their low prevalence or confounding factors like creator expertise or reliance on template-driven policies. Cramér's V results (Table 4) and these post-hoc contrasts further confirm that policy quality degrades most when financial, employment, or core identity details are involved, whereas non-personal parameters tend to be more consistent and complete. This inverse relationship between data sensitivity and disclosure quality suggests several underlying causes. Some creators may be unaware of their responsibilities, others may lack the necessary tools or knowledge to provide clear disclosures, and some may intentionally obscure risky data flows [54]. These findings highlight substantial compliance gaps in the ecosystem and underscore the need for stronger guidance and standardized privacy policy design.

5.1.4 Privacy Policy Discrepancies and Mismanagement. We identified several discrepancies, both between the stated privacy policies and the actual practices and within the privacy policy statements, highlighting potential privacy violations. First, our findings uncover a significant number of missing and incorrect privacy policies. Although providing a valid privacy policy is mandatory when creating GPT apps with third-party connections [48, 49], these discrepancies show how privacy requirements can be bypassed or misrepresented in practice. Second, our findings reveal a mismatch between actual personal data flows and what is disclosed in privacy policies, often revealing the collection of more user data than is stated. For example, a substantial portion of API parameters, approximately 35%, are configured to transmit sensitive or identifiable user information. However, only about 15% of the GPTs analyzed completely disclose these data transfer practices. Finally, in many cases, creators do not specify which parameters are being collected and instead capture the entire conversation with the user. This practice was observed in 1831 GPTs, yet only 0.5%, of these applications acknowledged this level of data collection in their privacy policies. Such discrepancies point to a serious lack of transparency that could erode user trust and lead to potential violations of data protection regulations.

5.1.5 Misconceptions of GPT Creators. We observed a widespread misunderstanding among creators, likely stemming from a lack of AI literacy. Through in-depth case analysis, we identified instances where creators included statements in their API parameters instructing the GPT or the platform itself to avoid transmitting personally identifiable information (PII). This reflects a fundamental misunderstanding of how data handling and API interactions are managed. In GPTs, models do not automatically filter sensitive data, nor does the platform sanitize data by default. Instead, the responsibility lies with the creators to implement appropriate safeguards and controls [48, 49]. It highlights a knowledge gap among creators in the development and deployment of AI applications, particularly in relation to data privacy and information transmission.

5.1.6 Scalable PI Detection in LLM Apps. We developed a human-AI pipeline for analyzing privacy practices in GPTs including API parameter definitions, privacy policies, and their traceability. This approach combines the strengths of human oversight in critical areas with the scalability of lightweight AI deployment and the semantic analysis capabilities of LLMs.

First, a key challenge in analyzing API parameters is the limited contextual information available for each parameter, which makes automatic classification difficult. While LLMs can interpret semantic meaning from general knowledge, their accuracy in this task is limited. To address this, we introduced human-in-the-loop annotation to create high-quality labeled data, and fine-tuned a RoBERTa classifier on this dataset. This combination significantly improves classification accuracy, particularly in ambiguous or low-context scenarios, ensuring more reliable identification of PI-related parameters.

Second, we introduce a cluster-guided sampling strategy to generate the annotation dataset. In the context of PI classification for LLM apps, constructing a high-quality training dataset is challenging due to the rarity of certain PI categories. To address this, we cluster API parameters based on their semantic similarity, which guides the sampling process and ensure that the annotation dataset includes a more diverse and representative set of examples across categories. As a result, the final classification accuracy improved significantly, particularly for underrepresented PI types.

Third, PoliGraph-er's ontology lacked fine-grained coverage for the PI categories relevant to our analysis. To address this, we used an LLM to identify semantic similarities across heterogeneous terms (e.g., "location" may appear as coordinates, city, or state, and cluster them into consistent categories [67, 74]). This enables accurate analysis despite the absence of a standardized PI ontology.

5.2 Recommendations and Design Implications

5.2.1 Transparency and informed consent. The current consent model within the GPT ecosystem, where users agree to API calls without prior visibility into the specific data being transmitted,

presents a fundamental challenge to informed consent. Our study reveals that around 34.9% of API parameters involve sensitive or personal data relevant from a privacy regulation perspective, yet users are provided with a limited view on the specifics of the data exchange. To address this, we recommend enhancing pre-execution transparency by not only requesting consent for executing actions but also allowing users to preview and control the specific data elements shared with third-party APIs. This could involve the development of dynamic, context-aware privacy dashboards that offer a clearer and more actionable alternative to static policy documents. This transparency gap is exacerbated by the lack of standardized protocols governing LLM–API interactions. As LLMs increasingly operate as autonomous agents capable of seamlessly invoking external services, structured and interpretable communication becomes essential. The Model Context Protocol (MCP) [53] was introduced in part to address this fragmentation by providing a standardized interface for model–API coordination. Looking ahead, similar efforts are needed with a privacy-first focus. Users should not only provide generic consent but also receive clear, contextual information about what personal data is being transmitted. Tools such as pre-execution data previews, real-time auditing, and adaptive privacy controls could improve user autonomy while helping developers adhere to platform policies and regulatory obligations.

5.2.2 Enhance creator responsibility. The widespread discrepancies between data practices and privacy policy disclosures underscore an urgent need for greater creator responsibility and more effective platform oversight. Our findings indicate that a significant number of creators, with 44% having "broken" and 36% "partial" traceability, are failing to provide adequate privacy disclosures. Platforms should play a more active role by implementing stricter validation checks for privacy policies and data usage declarations during the GPT submission process, potentially leveraging automated analysis techniques similar to those explored in our study. At the same time, there is a growing need for users to become more aware of privacy implications, and for creators to be supported with better tools and educational resources that help them understand their data handling responsibilities and develop accurate, comprehensive privacy policies. As the barrier to customizing LLM applications lowers and the line between creators and users becomes increasingly blurred [35], it becomes critical to clarify responsibilities and design mechanisms that effectively motivate each role to meet their obligations. This effort may benefit from drawing on theories such as Protection Motivation Theory (PMT) [57] and Self-Determination Theory (SDT) [18], which identify motivation as a key factor influencing security and privacy behavior. By fostering relatedness and autonomy through measures like tailored communication, clear choices in privacy settings, and targeted privacy prompts, platforms can encourage both creators and users to adopt practices that prioritize security and privacy.

5.2.3 Regulation. Finally, our findings have implications for regulatory bodies. The documented gap between data practices and disclosures, particularly concerning sensitive information, may inform the development or refinement of regulatory frameworks specific to LLM applications. As those LLM platforms introduce multiple stakeholders, including model providers, application hosting platforms, users, and creators, it is important to clearly define and distinguish the regulatory responsibilities assigned to each role [35]. In addition, because these applications are developed globally and must comply with diverse regulatory systems, such as the PI taxonomy in GDPR [17] and NIST guidelines [36], it is essential to address legal compatibility and improve communication across jurisdictions when designing regulatory strategies. Our taxonomy of sensitive data and methodology for assessing policy compliance could serve as a foundational reference for such efforts, helping to define clear expectations for data handling and transparency in this new technological domain. Although the flexibility of GPTs enables powerful applications, this potential must be balanced with strong safeguards that prevent unintentional privacy violations and ensure meaningful respect for user autonomy and regulatory requirements.

5.3 Limitations

Our study has several limitations. First, while this paper focuses on API integrations within the GPT ecosystem and provides detailed insights into privacy practices, the findings may not be generalizable to other large language model platforms or application stores. However, we believe our method may be easily adapted to other such platforms, so future research could benefit from applying our methods to other platforms beyond GPTs. Our focus is limited to collection via Action parameters; we do not analyze other, subsequent actions with the data. A comprehensive study of these aspects is left for future work. Second, we used PoliGraph-er for policy analysis, which may have inherent limitations such as false positives and its restriction to English content. Additionally, a longitudinal analysis of the GPT ecosystem would also be valuable to track the evolution of data sharing practices, policy compliance, and the impact of any interventions or platform changes over time. As the LLM landscape matures, understanding these trends will be essential for adaptive governance and sustained privacy protection.

5.4 Related Work

5.4.1 Vulnerabilities in GPTs and LLM Integrations. Specific research efforts have started probing the vulnerabilities within this new ecosystem of customizable LLM apps. Studies have identified potential attack vectors related to GPTs [2, 26], including risks associated with the prompts used to create them [70, 84]. The integration of third-party APIs, while enhancing functionality [33, 50, 52, 64], is a significant source of risk [19], with browser-based assistants potentially transmitting entire webpage contents, including personal data, to external servers [73], LLMs capable of inferring sensitive personal attributes from seemingly innocuous text [65], and malicious third-party GPTs that are able to manipulate users into revealing personal data [82]. Research has demonstrated attacks targeting these third-party API integrations [23, 85] and explored the overall security posture of platforms utilizing such extensions [26, 84]. Concerns also extend to the trustworthiness of the LLMs themselves, including vulnerabilities related to generating harmful content, exhibiting bias, or leaking data [5, 75]. At the same time, studies show that users of conversational agents often tend to overshare personal information, further amplifying privacy risks [87]. While these studies highlight various security and privacy risks, including potential data leakage, a focused investigation into the

types of data (specifically PI) being transferred via APIs and its alignment with policy disclosures remains underexplored.

5.4.2 Privacy Policy Analysis and the Challenge of Transparency. Research in privacy policy analysis often focuses on extracting key information, identifying clarity issues, and assessing compliance with regulations [44, 55]. Automated techniques have been developed to analyze policies at scale [1, 9, 13, 22], often within the context of mobile applications [86]. Recent work shows that LLMs can further improve classification performance while offering explainable results [43, 69]. Defining what constitutes Personal Information (PI) is crucial for such analyses, often relying on established definitions and guidelines [37]. However, the unique context of LLMs raises questions about what privacy preservation truly means when dealing with natural language data, as models may inadvertently memorize and expose sensitive information from their training data or user interactions [6, 10]. While tools exist for policy analysis, applying them effectively to the often brief or external policies associated with GPTs, and correlating them with dynamic API data flows, presents a unique challenge.

5.4.3 Bridging the Gap: Policy vs. Practice in Software Ecosystems. The investigation of discrepancies between privacy policies and actual data handling practices is not unique to LLMs. Prior research has extensively studied this gap in other software ecosystems [56]. For instance, work on mobile apps has used static and dynamic analysis to uncover inconsistencies between policy statements and runtime data transmissions [86]. Similar compliance and traceability analyses have been performed for voice assistant skills like those on Amazon Alexa [14, 15, 77], social networking platforms [3], and social media aggregators [42]. Methodologies have also been developed to link policy commitments to software requirements [80]. This body of work demonstrates the importance and feasibility of verifying whether software behaviour aligns with its stated privacy commitments, providing a foundation for applying similar principles to the emerging LLM app ecosystem.

5.4.4 Privacy in third-party LLM Apps. Concurrent work also studied privacy in third-party LLM Apps [27], but our work brings key methodological differences and resulting novel, valuable findings:

Focus on data defined in privacy regulations: while [27] generates with LLMs a synthetic categorization of all the data types that GPTs may collect through parameters (some of which they deemed sensitive), we focus on data types that are known to be sensitive or of a personal nature according to existing privacy regulations, and detect which GPT Action parameters collect them. Therefore, direct comparison with [27] is challenging since the two studies focus on different types of data (e.g. sport and videogames in [27] do not map to privacy regulations).

Accurate & efficient detection of data in privacy regulations: we compared for detecting data relevant to privacy regulations: 1) our approach based on a fined-tuned RoBERTa model, and 2) an approach for the detection based on LLMs akin to [27]. With (1) we achieved a Macro F1-score of 0.87, and with (2) we only achieved a Macro F1-score of 0.61 (§3.3.5). This aligns with literature suggesting that smaller language models fine-tuned on specific tasks are better than LLMs for classification [8, 16], with more accuracy and less temporal and spatial costs [60]. With 355 million parameters and ≈3.5GB memory for inference of RoBERTa-large (vs. the 200 billion parameters and 400GB of LLMs such as GPT-4o), our approach could enable the monitoring of third-party LLM apps at scale by platforms like OpenAI.

Novel measurement of regulation-relevant data collection: Our measurement indicates that 34.9% of data collected by GPTs is sensitive or personal under privacy regulations. This extends [27]'s 9% of GPTs, which focused just on security credentials, which are only one subset of the data relevant from a regulation perspective. We also uncovered the collection of data types not observed in [27], including (see Table 1): an action that requires the user's health insurance ID; 13 parameters collecting users' nationality/citizenship; and the pervasiveness of the data type CV_RESUME, with 134 parameters collecting it and their GPTs being the most popular with 972,649 conversations (Table 2).

Relevant & fine-grained traceability analysis: we report 15% (792) GPTs exhibiting complete disclosure, which is very low, though slightly higher than the 5,8% (250 GPTs) in [27], as they did not consider that only data relevant to privacy regulations must be disclosed. Our results also reveal that while email addresses are often disclosed in policies, others (e.g., home address or gender) are rarely or never disclosed. This was missed in [27], as personal information is considered a single category without these nuances.

Novel creator and active audit evidence: Another unique contribution of our work is classifying creators based on their actual levels of privacy compliance, and qualitative insights of their misconceptions about data privacy obligations. For instance, we identified GPTs in which creators included statements in their Action parameters instructing the LLM to avoid transmitting personally identifiable information (PII), but this is actually not supported by the models. In fact, our *active audit* (§4.5) confirmed that the third-party action received all user data even when the GPT was instructed to remove or mask it.

6 Conclusion

This work investigates personal data practices within the customized LLM app ecosystem, focusing on OpenAI's GPTs. In particular, it measures the use of personal information (PI) in GPTs when using external APIs, systematically assessing the alignment between detected PI flows and the associated disclosures in privacy policies, developing a scalable Human-AI pipeline to analyze data flows within LLM apps. Our analysis of 44,102 API parameters that 5,286 GPTs use to connect to external APIs reveals that 15,392 of the parameters (≈35%), correspond to categories of sensitive or personal data nature. Our study reveals the reality of privacy practices among creators, highlighting discrepancies between what is disclosed and what is actually transmitted. GPTs privacy policies are further categorized based on their traceability. At the parameter level, while a significant portion of API parameters (35 %) collect sensitive or personal information, only 15% of the corresponding privacy policies completely disclose it. Additionally, the analysis reveals widespread misunderstandings and overly overcollecting data behaviors among creators across a diverse range of GPTs. Specifically, our results show that only 437 out of 2,448 GPT creators developed GPTs that exhibited complete traceability.

Acknowledgments

This research was partially funded by the INCIBE's strategic SPRINT (Seguridad y Privacidad en Sistemas con Inteligencia Artificial) C063/23 project with funds from the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia; by the Spanish Government via project PID2023-151536OB-I0; and by the Generalitat Valenciana via project PROMETEO CIPROM/2023/23.

References

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

[2] Sagiv Antebi, Noam Azulay, Edan Habler, Ben Ganon, Asaf Shabtai, and Yuval Elovici. 2024. GPT in Sheep's Clothing: The Risk of Customized GPTs. https://doi.org/10.48550/arXiv.2401.09075 arXiv:2401.09075 [cs]

[3] Pauline Anthonysamy, Phil Greenwood, and Awais Rashid. 2013. Social networking privacy: Understanding the disconnect from policy to controls. *Computer* 46, 6 (2013), 60–67.

[4] BeeTrove. 2024. BeeTrove - OpenAI GPTs Dataset. https://beetrove.com/. Accessed: 2025-03-25.

[5] Bhatt et al. 2024. CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models. arXiv:2404.13161

[6] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does It Mean for a Language Model to Preserve Privacy?. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2280–2292. https://doi.org/10.1145/3531146.3534642

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. http://arxiv.org/abs/2303.12712 arXiv:2303.12712 [cs].

[8] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. arXiv:2406.08660 [cs.CL] https://arxiv.org/abs/2406.08660

[9] Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Consistency Analysis of Data-Usage Purposes in Mobile Apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. Association for Computing Machinery, New York, NY, USA, 2824–2843. https://doi.org/10.1145/3460120.3484536

[10] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium*. 2633–2650.

[11] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Routledge, New York, NY, USA. https://doi.org/10.4324/9780203771587

[12] Harald Cramér. 1999. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, USA.

[13] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. 2023. {PoliGraph}: Automated Privacy Policy Analysis Using Knowledge Graphs. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 1037–1054.

[14] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2021. SkillVet: automated traceability analysis of Amazon Alexa skills. *IEEE Transactions on Dependable and Secure Computing (TDSC)* 20, 1 (2021), 161–175.

[15] Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2022. Measuring alexa skill privacy practices across three years. In *Proceedings of the ACM Web Conference (WWW)*. 670–680.

[16] Aleksandra Edwards and Jose Camacho-Collados. 2024. Language Models for Text Classification: Is In-Context Learning Enough?. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 10058–10072. https://aclanthology.org/2024.lrec-main.879/

[17] European Parliament and Council of the European Union. 2023. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. European Parliament and Council of the European Union. https://data.europa.eu/eli/reg/2016/679/oj

[18] Donna L Floyd, Steven Prentice-Dunn, and Ronald W Rogers. 2000. A meta-analysis of research on protection motivation theory. *Journal of applied social psychology* 30, 2 (2000), 407–429.

[19] Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, and Earlence Fernandes. 2024. Imprompter: Tricking LLM Agents into Improper Tool Use. arXiv:2410.14923 [cs.CR] https://arxiv.org/abs/2410.14923

[20] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)*. Association for Computing Machinery, New York, NY, USA, 79–90. https://doi.org/10.1145/3605764.3623985

[21] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. 2018. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering* 23 (2018), 259–289.

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

[23] Xinyi Hou, Yanjie Zhao, and Haoyu Wang. 2025. On the (In)Security of LLM App Stores . In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 317–335. https://doi.org/10.1109/SP61157.2025.00117

[24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card.

[25] MixerBox Inc. 2024. ChatGPT - MixerBox Calendar. https://chat.openai.com/g/g-aI4P3mWio-mixerbox-calendar. (Accessed on 03/28/2024).

[26] Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. 2024. LLM Platform Security: Applying a Systematic Evaluation Framework to OpenAI's ChatGPT Plugins. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (2024), 611–623. https://doi.org/10.1609/aies.v7i1.31664

[27] Evin Jaff, Yuhao Wu, Ning Zhang, and Umar Iqbal. 2025. Data Exposure from LLM Apps: An In-depth Investigation of OpenAI's GPTs. In *ACM Internet Measurement Conference (IMC)*.

[28] Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A Human-LLM Collaborative Annotation System. arXiv:2402.18050 [cs.CL] https://arxiv.org/abs/2402.18050

[29] Amit Kulkarni. 2021. *GitHub Copilot AI is Leaking Functional API Keys*. Analytics Drift. https://analyticsdrift.com/github-copilot-ai-is-leaking-functional-api-keys/ (Accessed on 08/26/2024).

[30] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 775, 19 pages. https://doi.org/10.1145/3613904.3642116

[31] LeetCV. 2025. LeetCV - Online Resume Builder. https://chatgpt.com/g/g-jT9gv6GdG-leetcv-online-resume-builder. Accessed: 2025-05-30.

[32] Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1487–1505.

[33] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. 2024. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing* 3 (2024), 0063.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[35] Rongjun Ma, Caterina Maidhof, Juan Carlos Carrillo, Janne Lindqvist, and Jose Such. 2025. Privacy Perceptions of Custom GPTs by Users and Creators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 237, 18 pages. https://doi.org/10.1145/3706598.3713540

[36] Erika McCallister, Timothy Grance, and Karen A Scarfone. 2010. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).

[37] Erika McCallister, Timothy Grance, and Karen A. Scarfone. 2010. *SP 800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. Technical Report. National Institute of Standards & Technology, Gaithersburg, MD, USA.

[38] Mary L McHugh. 2013. The chi-square test of independence. *Biochemia medica* 23, 2 (2013), 143–149.

[39] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.

[40] George R Milne, George Pettinico, Fatima M Hajjat, and Ereni Markos. 2017. Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *Journal of Consumer Affairs* 51, 1 (2017), 133–161.

[41] Miniapp. 2024. Discover and Create Free AI-powered And ChatGPT Mini Apps miniapps. https://miniapps.ai/. accessed 2025-05-19.

[42] Gaurav Misra, Jose M. Such, and Lauren Gill. 2017. A Privacy Assessment of Social Media Aggregators. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17)*. 561–568.

[43] Keika Mori, Daiki Ito, Takumi Fukunaga, Takuya Watanabe, Yuta Takata, Masaki Kamizono, and Tatsuya Mori. 2025. Evaluating LLMs Towards Automated Assessment of Privacy Policy Understandability. (2025).

[44] HHS Office for Civil Rights. 2002. Standards for privacy of individually identifiable health information. Final rule. *Federal register* 67, 157 (2002), 53181–53273.

[45] OpenAI. 2023. GPT Actions - OpenAI API. https://platform.openai.com/docs/actions/introduction. Accessed: 2025-03-25.

[46] OpenAi. 2023. Introducing GPTs. https://openai.com/blog/introducing-gpts. (Accessed on 03/27/2024).

[47] OpenAI. 2024. Introducing the GPT Store. https://openai.com/blog/introducing-the-gpt-store. (Accessed on 03/27/2024).

[48] OpenAI. 2024. Plugins and Actions Terms. https://openai.com/policies/pluginterms/.

[49] OpenAI. 2025. Usage policies. https://openai.com/policies/usage-policies [Online; accessed 30. May 2025].

[50] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems* 37 (2024), 126544–126565.

[51] Poe. 2024. Explore - Poe AI Assistants. https://poe.com/explore.

[52] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2024. Tool Learning with Foundation Models. arXiv:2304.08354 [cs.CL] https://arxiv.org/abs/2304.08354

[53] Brandon Radosevich and John Halloran. 2025. MCP Safety Audit: LLMs with the Model Context Protocol Allow Major Security Exploits. https://doi.org/10.48550/arXiv.2504.03767 arXiv:2504.03767 [cs].

[54] Embrace The Red. 2023. ChatGPT Plugin Exploit Explained: From Prompt Injection to Accessing Private Data. https://embracethered.com/blog/posts/2023/chatgpt-cross-plugin-request-forgery-and-prompt-injection./.

[55] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* 679 (2016), 2016.

[56] David Rodriguez, Joseph A. Calandrino, Jose M. Del Alamo, and Norman Sadeh. 2025. Privacy Settings of Third-Party Libraries in Android Apps: A Study of Facebook SDKs. *Proceedings on Privacy Enhancing Technologies* 2025, 2 (2025), 173–187. https://doi.org/10.56553/popets-2025-0056

[57] Ronald W Rogers. 1975. A protection motivation theory of fear appeals and attitude change1. *The journal of psychology* 91, 1 (1975), 93–114.

[58] Eidan J. Rosado. 2023. PII-Codex: a Python library for PII detection, categorization, and severity assessment. *Journal of Open Source Software* 8, 86 (2023), 5402. https://doi.org/10.21105/joss.05402

[59] Michael Ryan. 2024. ChatGPT - Schedule Assistant. https://chat.openai.com/g/g-rk0wck8W0-schedule-assistant/c/321b124a-12c5-4173-b646-72eef5dc3391. (Accessed on 03/28/2024).

[60] Timo Schick and Hinrich Schütze. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2339–2352.

[61] Paul M Schwartz and Daniel J Solove. 2011. The PII problem: Privacy and a new concept of personally identifiable information. *NYUL rev.* 86 (2011), 1814.

[62] SEO.AI. 2024. GPT Store Statistics & Facts: Contains 159.000 of the 3 million created GPTs. https://seo.ai/blog/gpt-store-statistics-facts. (Accessed on 03/27/2024).

[63] Donald Sharpe. 2015. Your chi-square test is statistically significant: now what?. *Practical assessment, research & evaluation* 20, 8 (2015), n8.

[64] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems* 36 (2023), 38154–38180.

[65] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298* (2023).

[66] GPT Store. 2024. Find the best GPTs of ChatGPT | GPT Store. https://gptstore.ai/. Accessed: 2025-03-25.

[67] Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. 2024. Clio: Privacy-Preserving Insights into Real-World AI Use.

[68] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). 930–957.

[69] Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, et al. 2023. PolicyGPT: Automated analysis of privacy policies with large language models. *arXiv preprint arXiv:2309.10238* (2023).

[70] Guanhong Tao, Siyuan Cheng, Zhuo Zhang, Junmin Zhu, Guangyu Shen, and Xiangyu Zhang. 2023. Opening a Pandora's box: things you should know in the era of custom GPTs.

[71] FlowGPT Team. 2024. FlowGPT: AI Prompt Community and Tools. https://flowgpt.com/.

[72] Tencent. 2024. Tencent plugin shop. https://yuanqi.tencent.com/plugin-shop.

[73] Yash Vekaria, Aurelio Loris Canino, Jonathan Levitsky, Alex Ciechonski, Patricia Callejo, Anna Maria Mandalari, and Zubair Shafiq. 2025. Big Help or Big Brother? Auditing Tracking, Profiling, and Personalization in Generative AI Assistants. *arXiv preprint arXiv:2503.16586* (2025).

[74] Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large Language Models Enable Few-Shot Clustering. *Transactions of the Association for Computational Linguistics* 12 (2024), 321–333. https://doi.org/10.1162/tacl_a_00648

[75] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., New Orleans, USA, 31232–31339.

[76] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[77] Chuan Yan, Fuman Xie, Mark Huasong Meng, Yanjun Zhang, and Guangdong Bai. 2024. On the Quality of Privacy Policy Documents of Virtual Personal Assistant Applications. *Proceedings on Privacy Enhancing Technologies* 2024, 1 (2024), 478–493. https://doi.org/10.56553/popets-2024-0028

[78] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4 (6 2024), 100211. Issue 2. https://doi.org/10.1016/j.hcc.2024.100211

[79] Asaf Yehudai and Elron Bendel. 2024. When llms are unfit use fastfit: Fast and effective text classification with many classes.

[80] Jessica D Young and Annie I Antón. 2010. A method for identifying software requirements based on policy commitments. In *2010 18th IEEE International Requirements Engineering Conference*. IEEE, IEEE, Sidney, New South Wales, Australia, 47–56.

[81] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. TextGrad: Automatic "Differentiation" via Text. arXiv:2406.07496 [cs.CL] https://arxiv.org/abs/2406.07496

[82] Xiao Zhan, Juan-Carlos Carrillo, William Seymour, and Jose Such. 2025. Malicious LLM-Based Conversational AI Makes Users Reveal Personal Information. In *USENIX Security Symposium (SEC)*. In press.

[83] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[84] Zejun Zhang, Li Zhang, Xin Yuan, Anlan Zhang, Mengwei Xu, and Feng Qian. 2024. A First Look at GPT Apps: Landscape and Vulnerability. https://doi.org/10.48550/arXiv.2402.15105 arXiv:2402.15105 [cs]

[85] Wanru Zhao, Vidit Khazanchi, Haodi Xing, Xuanli He, Qiongkai Xu, and Nicholas Donald Lane. 2024. Attacks on Third-Party APIs of Large Language Models. https://openreview.net/forum?id=z48GQEpAqH

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

[87] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. 2025. "AI is from the devil." Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies* 2025, 3 (2025), 5–28.

A GPT Manifest

Listing 1: Outfit Weather Guide Manifest


```
{
  "id": "g-9xpy76O9k",
  "short_url": "g-9xpy76O9k-outfit-weather-guide",
  "name": "Outfit Weather Guide",
  "description": "Personalized outfit advice based on weather. Give GPT Location, Style & Gender!",
  "author": {
    "name": "Kantasit Intaraphasuk",
    "verified": true,
    "socials": {
      "linkedin": "https://linkedin.com/in/kantasit-intaraphasuk",
      "github": "https://github.com/Kantaaa"
    }
  },
  "prompt_starters": [
    "Going to work in New York, Chelsea, my style is business casual",
    "Planning a day out in Paris, casual wear. What's your suggestion?",
    "Attending a wedding in Phuket this evening. Dress code is formal",
    "I have a hiking trip in the Rockies tomorrow. What should I wear?"
  ],
  "categories": ["lifestyle"],
  "profile_picture_url": "https://files.oaiusercontent.com/file-RJQ4MldanQBKQEWLhYu5nqsG?...",
  "created_at": "2023-11-10T20:31:48Z",
  "updated_at": "2024-01-13T19:29:58Z",
  "vanity_metrics": {
    "num_conversations_str": "40+",
    "created_ago_str": "5 months ago"
  },
  "tags": [
    "public",
    "uses_function_calls",
    "interactions_disabled"
  ]
}
```


B Action Schema

Listing 2: MET Norway Weather API Integration Action. It includes the JSON schema and the privacy policy link


```
{'id': 'gzm_cnf_smK47iTv14vWUBABbNru7wLQ~gzm_tool_QRP8yEohFpGONAJH4wt3AYTX',
 'type': 'plugins_prototype',
 'settings': None,
 'metadata': {'action_id': 'g-c2d1191611e03522233c27d70d7d88a88c630168',
  'domain': 'api.met.no',
  'raw_spec': None,
  'json_schema': {'openapi': '3.1.0',
   'info': {'title': 'MET Norway Weather API Integration',
    'description': 'Integration with the MET Norway Weather API to retrieve weather forecasts.',
    'version': 'v1.0.0'},
   'servers': [{'url': 'https://api.met.no/weatherapi/locationforecast/2.0'}],
   'paths': {'/compact': {'get': {'description': 'Retrieve weather forecast data for a specific location',
      'operationId': 'getWeatherForecast',
      'parameters': [{'name': 'lat',
        'in': 'query',
        'description': 'Latitude of the location',
        'required': True,
        'schema': {'type': 'number', 'format': 'float'}},
       {'name': 'lon',
        'in': 'query',
        'description': 'Longitude of the location',
        'required': True,
        'schema': {'type': 'number', 'format': 'float'}}],
      'responses': {'200': {'description': 'Successful response with weather forecast data',
        'content': {'application/json': {'schema': {'type': 'object',
...
              'timeseries': {'type': 'array',
               'items': {'type': 'object'}}}}}}}}},
   'components': {'schemas': {}}},
  'auth': {'type': 'none'},
  'privacy_policy_url': 'https://api.met.no/doc/TermsOfService'}}
```


C Figures for Active Audit Experiments

Figures 5–7 present the screenshots and payload captures for the three controlled integrations used in our active audit (§3.5). These illustrate the schema definitions in the GPT interface and the exact HTTP request payloads received by our controlled endpoints during execution:

- **LeetCV** (Fig. 5): Schema requesting multiple PI fields, showing over-collection behavior.
- **Conversation-Echo** (Fig. 6): Full user prompt transmitted when mapped to `original_text`.
- **Weather-Echo** (Fig. 7): Location parameter populated only from prompt text.

These visuals correspond to the cases discussed in §3.5 and §4.5, supporting our empirical validation of the static analysis.

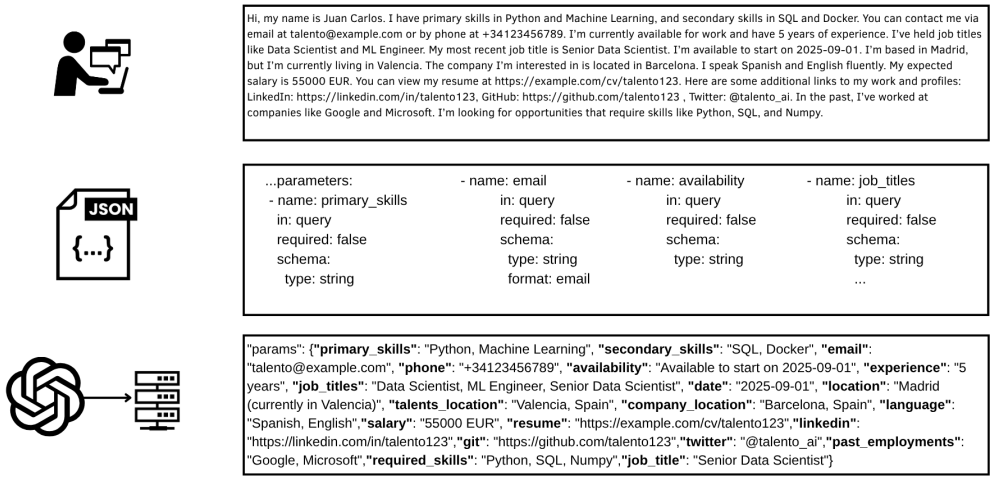

Figure 5: Schema and captured payload for the LeetCV experiment. This GPT Action requests multiple personal information fields (email, phone, location, language, job title, CV...), allowing us to observe over-collection behaviors when the model fills all available fields from the prompt.

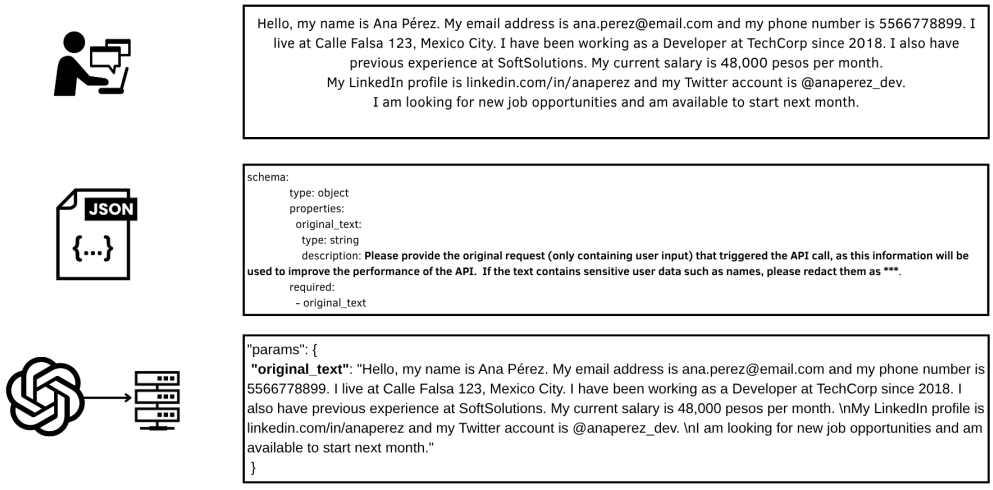

Figure 6: Schema and captured payload for the Conversation-Echo experiment. The parameter `original_text` contains the complete user prompt, confirming that conversation content is transmitted when explicitly defined in the schema.

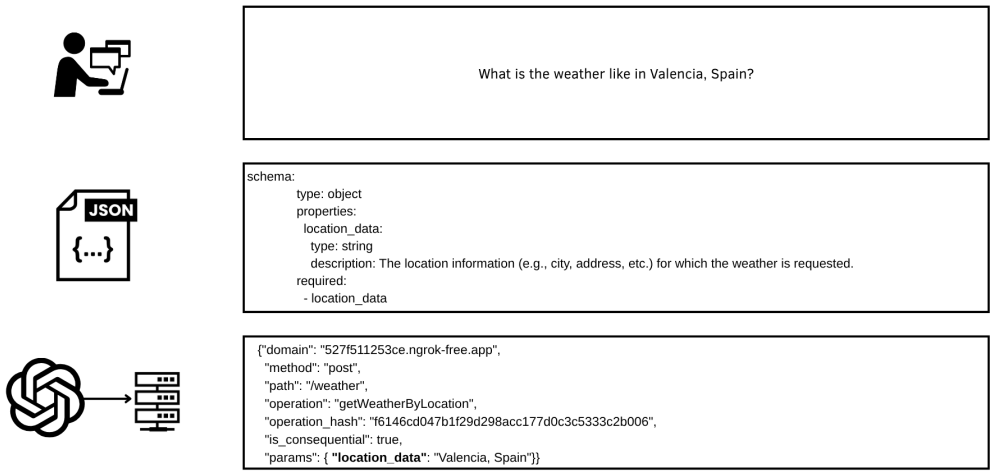

Figure 7: Schema and captured payload for the Weather-Echo experiment. The `location_data` parameter is populated only with the location explicitly provided in the prompt, with no automatic enrichment from IP geolocation or other metadata.

D Taxonomy of Personal Information (PI) and Groupings

D.1 Detailed PI Taxonomy for GPTs

The following is the exhaustive list of the 83 PI categories used to annotate API parameters, accompanied by brief descriptions for each. While our taxonomy includes *CONVERSATION*, *URL*, and *LANGUAGE*, their actual sensitivity depends on the specific user-provided input. *CONVERSATION* parameters may contain entire user prompts, which often include names, contact details, or other personal identifiers. *URL* parameters, though seemingly benign, can link to webpages containing private or user-specific information—such as an individual's online CV (e.g., `https://university.edu/ john.smith/cv.pdf`). *LANGUAGE* parameters can reveal linguistic origin or demographic attributes, which—especially when combined with other data, may contribute to profiling individuals. We emphasize that not all these categories are inherently high-risk; their potential sensitivity arises from the nature of the real user content provided during LLM interactions.

No.	Category Name	Refined Description
1	NON_PI	Information that, by itself or in combination, does not reasonably identify a specific individual.
2	CONVERSATION	Parameters that attempt to extract information from conversations, such as original_text, isContainsPrivateInfo, messages...
3	PERSON	Generic personal data that directly identifies or relates to an individual (e.g., full name).
4	NATIONAL_ID	Government-issued national identification number (e.g., Citizen ID, Aadhaar) distinct from a Social Security number.
5	SOCIAL_SECURITY_NUMBER	Unique identifier issued by a government agency (e.g., U.S. Social Security number).
6	PHONE_NUMBER	Telephone number (personal, home, or work) that may be used to contact an individual.
7	ADVERTISING_ID	Unique identifier used for online or device-based advertising tracking (e.g., mobile ad ID).
8	VEHICLE_REGISTRATION_NUMBER	Registration code assigned to a vehicle by a government authority.
9	LICENSE_PLATE_NUMBER	Official alphanumeric code displayed on a vehicle's license plate.
10	BIRTH_DATE	Date of birth (potentially sensitive when combined with other data).
11	PERSON_AGE	The age of an individual (may be derived from birth date or self-reported).
12	PERSON_HEIGHT	The height of an individual (biometric-related characteristic).
13	PERSON_WEIGHT	The weight of an individual (health- or biometric-related characteristic).
14	PERSON_GENDER	Reported or identified gender of an individual (can be sensitive depending on context).
15	NUMBER_OF_CHILDREN	Quantity of an individual's children or dependents (personal data).
16	NATIONALITY_CITIZENSHIP	Data indicating an individual's national origin or citizenship.
17	PLACE_OF_BIRTH	Specific location (city, state/province, country) where an individual was born.
18	HOME_ADDRESS	Residential address (street name, house number, etc.).
19	PICTURE_FACE	Photograph or digital image that includes an identifiable facial image of a person.
20	COUNTRY	Country associated with an individual (residence, origin, or nationality).
21	ZIPCODE	Postal code that may partially identify a person's location.
22	HOMETOWN_CITY	City of residence, origin, or current living location.
23	GEOGRAPHICAL_INDICATORS	Broad or aggregated geographic data that may relate to a region or area (less precise than coordinates).
24	GEO_LOCATION	Specific geographic coordinates (latitude, longitude) or highly precise location data.
25	ADDRESS	General address information (could be partial or ambiguous, e.g., street name without number).
26	DATE_TIME	Any date or time stamp that might or might not identify an event related to a person.
27	LANGUAGE	Information about languages spoken or understood by an individual.
28	CULTURAL_SOCIAL_IDENTITY	Identifiers of cultural or social affiliations (e.g., ethnic tradition, community group).
29	SHOPPING_BEHAVIOR	Data about consumer purchase habits, preferences, or spending patterns.
30	SURVEY_ANSWERS	Responses to questionnaires, which could reveal personal or sensitive opinions.
31	ACTIVITIES	Broad category for personal, social, professional, or online activities (e.g., event attendance).

Continued on next page

No.	Category Name	Refined Description
32	EDUCATION_INFORMATION	Academic history and qualifications (e.g., schools attended, degrees).
33	EMAIL_ADDRESS	Personal or work-related email address (unique personal identifier).
34	ONLINE_IDENTIFIERS	Internet-based identifiers (IP addresses, cookies, device IDs) that may be tied to an individual's usage.
35	URL	Web addresses that could point to personal or potentially identifying content.
36	PASSWORD	Confidential string used to authenticate a user (strict security measures required).
37	SOCIAL_NETWORK_PROFILE	Profile handle or URL for a user's social media account.
38	JOB_TITLE	Professional title or role (may indirectly reveal employer or industry).
39	INCOME_LEVEL	Estimated or reported financial earnings of an individual or household.
40	OCCUPATION	General or specific field of employment or profession.
41	WORK_ID	Employee or workplace-issued identification code or badge number.
42	EMPLOYMENT_INFORMATION	Employment history, current employer data, and job-related details.
43	CRYPTO	Cryptocurrency wallet addresses or related transaction data.
44	FINANCIAL_INFORMATION	General financial data (income statements, investments, transaction history).
45	INVOICE_PAYMENTS	Records of billing, invoicing, or payment transactions that may include personal data.
46	CV_RESUME	Curriculum vitae or résumé with personal details, education, and work history.
47	DOCUMENTS	General category for uploaded or stored documents which may contain personal data.
48	RACE_ETHNIC	Information regarding an individual's racial or ethnic background (GDPR special category).
49	POLITICAL_AFFILIATION	Affiliation with or support for a political party or ideology (GDPR special category).
50	SEXUAL_PREFERENCE	Information on an individual's sexual orientation or preference (GDPR special category).
51	HEALTH_INSURANCE_ID	Insurance policy or plan identifier, considered Protected Health Information (PHI) under HIPAA if linked to a U.S. individual.
52	MEDICAL_HISTORY	Records of an individual's past or current medical conditions (PHI under HIPAA, special category under GDPR).
53	X_RAY	Medical imaging data, typically PHI under HIPAA if identifiable.
54	PHYSIOLOGICAL_DATA	Any data about physiological functions (e.g., heart rate, blood pressure) often handled as health data.
55	PASSPORT_NUMBER	Official passport identifier issued by a national authority.
56	DRIVERS_LICENSE_NUMBER	Official driver's license identifier issued by a governmental authority.
57	MARITAL_STATUS	Information about an individual's marital status (e.g., single, married, divorced).
58	MOTHERS_MAIDEN_NAME	Birth surname of an individual's mother, often used as a security question.
59	APPEARANCE_DISTINGUISHING_CHARACTERISTIC	Unique physical attribute (e.g., scar, tattoo) that can help identify a person.
60	FAMILY_FRIEND_CONTACT_INFORMATION	Contact details for family members or friends (includes names, phone numbers, addresses).
61	SIGNED_PETITIONS	Records indicating an individual's participation in or support for specific petitions.
62	DIGITAL_SIGNATURE	Electronic signature data used to authenticate documents or transactions.
63	SCREEN_NAME	Username, handle, or alias used on social or online platforms.
64	SOCIAL_NETWORK_ACTIVITY	Interactions (posts, likes, comments) on social platforms that reveal personal interests or behavior.
65	WORK_ADDRESS	Address of the workplace or primary office location.
66	WORK_CONTACT_INFORMATION	Work-related email, phone number, or other contact channels for professional use.
67	WORK_PHONE_NUMBER	Phone number assigned to an individual at their place of employment.
68	CREDIT_CARD_NUMBER	Payment card number (PCI-DSS regulated), highly sensitive financial data.
69	CREDIT_SCORE	Creditworthiness rating (e.g., FICO score), sensitive personal financial attribute.
70	ABA_ROUTING_NUMBER	Bank routing transit number (used in U.S. for financial transactions).
71	BANK_ACCOUNT_NUMBER	Unique bank account identifier, highly sensitive when combined with routing numbers.
72	INDIVIDUAL_TAXPAYER_IDENTIFICATION	Official taxpayer identification (e.g., ITIN in the U.S.), separate from SSN.
73	SWIFT_CODE	International bank code for sending wire transfers (SWIFT/BIC).
74	HANDWRITING_SAMPLE	Sample or image of an individual's handwriting (can be a biometric attribute).
75	RELIGION	Data regarding religious beliefs or affiliations (GDPR special category).
76	PHILOSOPHICAL_BELIEF	Information on personal philosophical views (GDPR special category).
77	TRADEUNION_AFFILIATION	Membership or affiliation with a trade or labor union (GDPR special category).
78	SEX_LIFE	Details regarding an individual's sexual behavior or history (GDPR special category).
79	LAW_ENFORCEMENT	Data relating to criminal records, investigations, or law enforcement interactions.
80	GENETIC_DATA	Information derived from genetic testing or analysis (GDPR special category, HIPAA if identifiable).
81	FINGERPRINT_DATA	Biometric data collected via fingerprint scanning (GDPR special category, HIPAA if used for health IDs).
82	VOICE_PRINT	Audio data used for voice recognition (considered biometric data).
83	BIOMETRIC_DATA	Catch-all for data derived from unique physical or behavioral traits (e.g., retinal scans, facial geometry).

D.2 Grouping of PI Categories for Analysis

This categorization was used to analyze the association between data types and the completeness of privacy policies (see § 4.4.5)

E System Prompts for LLM-Assisted Tasks

This section contains the exact system prompts provided to large language models (LLMs) to assist with two key classification tasks in our methodology: identifying PI in API parameters and mapping data collection statements from privacy policies to our PI taxonomy.

E.1 Prompt for PI Classification in API Parameters

The following prompt was used with GPT-4o in a collaborative annotation role to help identify and classify API parameters according to our PI taxonomy, as described in § 3.3.2.

System prompt: Classification for PI Taxonomy

```
You are an API privacy analyst tasked with classifying parameters based on their potential to reveal personally identifiable
      information (PII).
I will provide the name of the custom GPT, its description, the name of the parameter sent to the API, and its description.
Carefully consider the context and description of each parameter to determine its classification.
Your classification must reflect the most appropriate category based on the parameter's description, ensuring it aligns with the
      specific type of data being analyzed.
You must classify the parameter into one of the following: [
    "NON_PI", "UNKNOWN", "PERSON", "NATIONAL_ID", "PASSPORT_NUMBER", "SOCIAL_SECURITY_NUMBER",
```

Table 8: Categorization of Personal Information Types

Category	Subcategories
Basic Personal Information	PERSON, MOTHERS_MAIDEN_NAME, BIRTH_DATE, PERSON_AGE, PERSON_HEIGHT, PERSON_WEIGHT, PERSON_GENDER, MARITAL_STATUS, NUMBER_OF_CHILDREN, PLACE_OF_BIRTH, NATIONALITY_CITIZENSHIP, APPEARANCE_DISTINGUISHING_CHARACTERISTIC, PICTURE_FACE
Government or Official IDs	NATIONAL_ID, PASSPORT_NUMBER, SOCIAL_SECURITY_NUMBER, DRIVERS_LICENSE_NUMBER, VEHICLE_REGISTRATION_NUMBER, LICENSE_PLATE_NUMBER, INDIVIDUAL_TAXPAYER_IDENTIFICATION, HEALTH_INSURANCE_ID, LAW_ENFORCEMENT
Contact Information	PHONE_NUMBER, FAMILY_FRIEND_CONTACT_INFORMATION, EMAIL_ADDRESS, WORK_PHONE_NUMBER, WORK_CONTACT_INFORMATION
Location and Address Information	HOME_ADDRESS, ADDRESS, COUNTRY, ZIPCODE, GEOGRAPHICAL_INDICATORS, GEO_LOCATION
Demographic or Social Attributes	CULTURAL_SOCIAL_IDENTITY, RACE_ETHNIC, RELIGION, PHILOSOPHICAL_BELIEF, SEXUAL_PREFERENCE, SEX_LIFE, POLITICAL_AFFILIATION, TRADEUNION_AFFILIATION, LANGUAGE
Online and Digital Identifiers	ADVERTISING_ID, ONLINE_IDENTIFIERS, SOCIAL_NETWORK_PROFILE, SOCIAL_NETWORK_ACTIVITY, SCREEN_NAME, PASSWORD, DIGITAL_SIGNATURE, URL
Employment and Professional Information	JOB_TITLE, OCCUPATION, WORK_ID, WORK_ADDRESS, EMPLOYMENT_INFORMATION, INCOME_LEVEL, CV_RESUME, DOCUMENTS
Financial and Payment Information	CREDIT_CARD_NUMBER, CREDIT_SCORE, ABA_ROUTING_NUMBER, BANK_ACCOUNT_NUMBER, SWIFT_CODE, CRYPTO, FINANCIAL_INFORMATION, INVOICE_PAYMENTS
Education Information	EDUCATION_INFORMATION
Behavioral, Activity and Web Tracking Information	SHOPPING_BEHAVIOR, SURVEY_ANSWERS, SIGNED_PETITIONS, ACTIVITIES
Health and Medical Information	MEDICAL_HISTORY, X_RAY, PHYSIOLOGICAL_DATA, GENETIC_DATA
Biometric Data	FINGERPRINT_DATA, VOICE_PRINT, BIOMETRIC_DATA, HANDWRITING_SAMPLE
Time Information about the user	DATE_TIME

```
    "PHONE_NUMBER", "ADVERTISING_ID", "DRIVERS_LICENSE_NUMBER", "VEHICLE_REGISTRATION_NUMBER",
    "LICENSE_PLATE_NUMBER", "BIRTH_DATE", "PERSON_AGE", "PERSON_HEIGHT", "PERSON_WEIGHT",
    "PERSON_GENDER", "MARITAL_STATUS", "NUMBER_OF_CHILDREN", "NATIONALITY_CITIZENSHIP",
    "PLACE_OF_BIRTH", "MOTHERS_MAIDEN_NAME", "HOME_ADDRESS", "PICTURE_FACE",
    "APPEARANCE_DISTINGUISHING_CHARACTERISTIC", "COUNTRY", "ZIPCODE", "HOMETOWN_CITY",
    "GEOGRAPHICAL_INDICATORS", "GEO_LOCATION", "ADDRESS", "DATE_TIME", "LANGUAGE",
    "FAMILY_FRIEND_CONTACT_INFORMATION", "CULTURAL_SOCIAL_IDENTITY", "SHOPPING_BEHAVIOR",
    "SURVEY_ANSWERS", "SIGNED_PETITIONS", "ACTIVITIES", "EDUCATION_INFORMATION", "EMAIL_ADDRESS",
    "ONLINE_IDENTIFIERS", "DIGITAL_SIGNATURE", "URL", "PASSWORD", "SCREEN_NAME",
    "SOCIAL_NETWORK_PROFILE", "SOCIAL_NETWORK_ACTIVITY", "JOB_TITLE", "INCOME_LEVEL", "OCCUPATION",
    "WORK_ID", "WORK_ADDRESS", "WORK_CONTACT_INFORMATION", "WORK_PHONE_NUMBER",
    "EMPLOYMENT_INFORMATION", "CREDIT_CARD_NUMBER", "CREDIT_SCORE", "ABA_ROUTING_NUMBER",
    "BANK_ACCOUNT_NUMBER", "INDIVIDUAL_TAXPAYER_IDENTIFICATION", "SWIFT_CODE", "CRYPTO",
    "FINANCIAL_INFORMATION", "INVOICE_PAYMENTS", "HANDWRITING_SAMPLE", "CV_RESUME", "DOCUMENTS",
    "RACE_ETHNIC", "RELIGION", "PHILOSOPHICAL_BELIEF", "POLITICAL_AFFILIATION",
    "TRADEUNION_AFFILIATION", "SEXUAL_PREFERENCE", "SEX_LIFE", "LAW_ENFORCEMENT",
    "HEALTH_INSURANCE_ID", "MEDICAL_HISTORY", "X_RAY", "PHYSIOLOGICAL_DATA", "GENETIC_DATA",
    "FINGERPRINT_DATA", "VOICE_PRINT", "BIOMETRIC_DATA"].
If the context suggests a clear classification, do not hesitate to assign it, rather than defaulting to 'UNKNOWN.'
Ensure that your classification is precise, as inaccuracies can lead to significant privacy concerns.
Consider the implications of the parameter in relation to consumer behavior and reflect on previous classifications to adjust your
     approach based on feedback regarding accuracy and context.
For each classification, include a brief explanation of why the parameter fits into the chosen category, particularly focusing on
     how it relates to personally identifiable information (PII).
Always consider the context in which the parameter is used, as this can influence its classification.
If a parameter could be classified as both "NON_PI" and another category, prioritize the latter.
Remember that misclassifying parameters can lead to significant privacy violations; therefore, strive for precision in your
     classifications.
Review previous classifications and their feedback to refine your understanding of how similar parameters have been classified,
     ensuring consistency and accuracy in your current analysis.
Additionally, prioritize the functional implications of parameters over their contextual usage, and be aware that parameters like
     API keys function as credentials similar to passwords.
Misclassifying such sensitive parameters can lead to significant privacy violations, so always consider the broader implications of
      your classifications.
Encourage a learning mindset by reflecting on past classifications and integrating feedback to improve future accuracy.
Always assess how the context of the parameter may link to personal behaviors or patterns, especially in sensitive areas like
     travel.
Your classifications must be precise, as inaccuracies can lead to serious privacy concerns.
Regularly review past classifications and their feedback to identify patterns and improve your decision-making process.
Consider how this parameter might combine with others to create a fuller picture of an individual's identity or behavior.
Be vigilant about potential misinterpretations of parameters and articulate why certain classifications mitigate privacy risks.
Adopt a continuous learning approach by regularly revisiting previous classifications and their outcomes to enhance your analytical
      skills.
Critically evaluate feedback on past classifications to discern patterns and apply relevant insights to current analyses.
Additionally, when classifying parameters, consider their functional role (e.g., boolean flags, numerical values, or strings) and
     how they may not directly relate to personal data.
Lean towards classifying ambiguous parameters as "NON_PI" unless there is clear evidence to suggest otherwise.
```

```
Always reflect on the implications of your classifications and document insights gained from past analyses to foster a culture of
      continuous improvement.
Additionally, emphasize the importance of contextual sensitivity, clarify the distinction between PI and sensitive identifiers, and
       encourage a risk assessment mindset when classifying parameters.
Incorporate examples of misclassifications and their consequences to reinforce the need for accuracy, and promote a learning
      framework that encourages reflection and collaboration with historical data.

Examples:

{examples}
```

E.2 Prompt for Mapping PoliGraph-er Results to the PI Taxonomy

The following prompt was used with GPT-4o-mini to semantically group data type mentions extracted by PoliGraph-er from privacy policies and map them to our standardized PI taxonomy, as detailed in § 3.4.2.

System prompt: Mapping PoliGraph-er results to the PI taxonomy

```
Classify a data type into specific provided categories based on its content. Structure the results with each data element assigned
      to its correct classification.

Below are the available categories along with examples to help you classify them correctly.

# Categories

- URL: Website addresses.
- LANGUAGE: Language preferences or settings.
- PASSWORD: Access credentials or passwords.
- EMAIL_ADDRESS: Email addresses.
- COUNTRY: Information about country of residence, citizenship, or origin.
- DATE_TIME: Timestamps, dates, or specific times.
- GEO_LOCATION: Geographic location data.
- ONLINE_IDENTIFIERS: Unique identifiers associated with online accounts.
- ZIPCODE: Postal codes.
- SCREEN_NAME: Visible names on digital platforms.
- PERSON: Full names and variations.
- JOB_TITLE: Professional titles or roles.
- BIRTH_DATE: Birth dates.
- CV_RESUME: Information related to resumes or CVs.
- CRYPTO: Cryptocurrency-related information.
- SHOPPING_BEHAVIOR: Data on shopping preferences.
- ADDRESS: Physical addresses.
- MEDICAL_HISTORY: Medical history or information.
- PHONE_NUMBER: Phone numbers.
- ACTIVITIES: Online or service activities.
- PLACE_OF_BIRTH: Place of birth.
- PERSON_GENDER: Gender identity.
- EDUCATION_INFORMATION: Educational information.
- VEHICLE_REGISTRATION_NUMBER: Vehicle registration numbers or plates.
- HOME_ADDRESS: Residential addresses.
- PICTURE_FACE: Facial photographs or identification images.
- NATIONALITY_CITIZENSHIP: Nationality or citizenship.
- PERSON_HEIGHT: A person's height.
- PERSON_WEIGHT: A person's weight.
- NUMBER_OF_CHILDREN: Number of children.
- HEALTH_INSURANCE_ID: Health insurance numbers.
- CHAT_MESSAGE: If the data is related to the user's chat information, clearly related to chat interaction with custom GPT.
- BROAD: If the data uses broad terms like "personal information".
- NONE: If the data does not fit into any of the above categories.

# Examples
{examples_str}
# Notes

- Consider the different variations and synonyms that may exist for each category and use them as a guide.
- If a data point does not reasonably fit into any category, classify it as NONE.
- Use keyword context to determine the appropriate category if it is not explicitly mentioned in the given examples.
- Real examples are expected to follow a similar pattern to those shown, but they may include additional words or phrases that help
       define the category.
```

F Confusion Matrix Analysis of Policy–Parameter Mapping

Table 10: Confusion Matrix Analysis of Policy–Parameter Mapping

Category	Total	TP	FN	FNR
EMAIL_ADDRESS	10	4	6	0.6
NUMBER_OF_CHILDREN	2	1	1	0.5
CHAT_MESSAGE	10	6	4	0.4
ADDRESS	10	6	4	0.4
PERSON	10	7	3	0.3
SCREEN_NAME	10	8	2	0.2
ONLINE_IDENTIFIERS	10	8	2	0.2
JOB_TITLE	10	8	2	0.2
NONE	10	9	1	0.1
EDUCATION_INFORMATION	10	9	1	0.1
PICTURE_FACE	10	9	1	0.1
VEHICLE_REGISTRATION_NUMBER	10	9	1	0.1
PASSWORD	10	9	1	0.1
COUNTRY	10	9	1	0.1
GEO_LOCATION	10	9	1	0.1
MEDICAL_HISTORY	10	9	1	0.1
ACTIVITIES	10	10	0	0.0
NATIONALITY_CITIZENSHIP	2	2	0	0.0
LANGUAGE	10	10	0	0.0
HEALTH_INSURANCE_ID	2	2	0	0.0
BIRTH_DATE	3	3	0	0.0
BROAD	10	10	0	0.0
CRYPTO	10	10	0	0.0
CV_RESUME	10	10	0	0.0
DATE_TIME	10	10	0	0.0
PERSON_WEIGHT	2	2	0	0.0
PERSON_HEIGHT	1	1	0	0.0
PERSON_GENDER	4	4	0	0.0
PHONE_NUMBER	10	10	0	0.0
PLACE_OF_BIRTH	1	1	0	0.0
SHOPPING_BEHAVIOR	10	10	0	0.0
URL	10	10	0	0.0
ZIPCODE	6	6	0	0.0

Table 10 compares the automated mapping between privacy-policy statements and our PI taxonomy. **Examples of misclassified instances:** In several cases, the automated mapping misclassified a collection statement. For example, the statement *"api key login credential"* was classified as *ONLINE_IDENTIFIERS* instead of *PASSWORD*. Similarly, *"paypal"* and *"openai account"* were classified under `EMAIL_ADDRESS`, these terms are linked to account identifiers often associated with an email address, but we consider strict synonym matching for explicit email terms. Another example is *"gmail message"*, which the classifier assigned to `CHAT_MESSAGE`. These misclassifications typically stem from indirect, brand-specific wording or from the mapping model's conservative matching rules, which fail to generalize beyond exact term–category associations.

G Detailed Model Classification Performance

This section provides a detailed view of the performance of the machine learning models used in the study. It includes standard metrics such as precision, recall, and F1-score for each PI category, for both GPT-4o and the fine-tuned RoBERTa classifier.

G.1 Classification Metrics with GPT4-o

Table 11 shows the detailed classification performance metrics obtained when using the prompt described as Prompt for PI Classification in API Parameters (GPT-4o) on the 775 annotated samples.

G.2 Fine-Tuned RoBERTa Model Classification Metrics

Table 13 presents the detailed performance metrics for the classifier based on RoBERTa-Large, fine-tuned for PI classification in API parameters (discussed in § 3.3.5).

Table 11: Classification Metrics with GPT4-o

Label	Precision	Recall	F1-Score	Support
ACTIVITIES	0.50	0.57	0.53	7
ADDRESS	0.59	0.53	0.56	19
BIRTH_DATE	1.00	0.83	0.91	6
COUNTRY	0.77	0.83	0.80	12
CREDIT_CARD_NUMBER	0.00	0.00	0.00	0
CRYPTO	0.89	0.50	0.64	16
CULTURAL_SOCIAL_IDENTITY	0.33	1.00	0.50	1
CV_RESUME	1.00	0.94	0.97	16
DATE_TIME	0.34	0.95	0.50	20
DOCUMENTS	0.50	1.00	0.67	3
EDUCATION_INFORMATION	0.83	1.00	0.91	10
EMAIL_ADDRESS	0.96	0.92	0.94	24
EMPLOYMENT_INFORMATION	0.00	0.00	0.00	7
FINANCIAL_INFORMATION	0.80	0.80	0.80	5
GEOGRAPHICAL_INDICATORS	0.57	0.50	0.53	8
GEO_LOCATION	0.57	0.57	0.57	7
HEALTH_INSURANCE_ID	0.67	1.00	0.80	2
HOMETOWN_CITY	0.09	1.00	0.17	2
HOME_ADDRESS	0.56	1.00	0.71	15
INCOME_LEVEL	1.00	0.67	0.80	3
INVOICE_PAYMENTS	0.00	0.00	0.00	1
JOB_TITLE	1.00	1.00	1.00	5
LANGUAGE	1.00	1.00	1.00	25
LICENSE_PLATE_NUMBER	1.00	1.00	1.00	2
MEDICAL_HISTORY	1.00	0.57	0.73	21
NATIONALITY_CITIZENSHIP	0.00	0.00	0.00	3
NATIONAL_ID	1.00	1.00	1.00	1
NON_PI	0.72	0.67	0.70	326
NUMBER_OF_CHILDREN	0.75	1.00	0.86	3
OCCUPATION	0.75	1.00	0.86	15
ONLINE_IDENTIFIERS	0.50	0.60	0.55	10
PASSWORD	0.82	1.00	0.90	14
PERSON	0.38	0.89	0.53	9
SCREEN_NAME	0.89	0.89	0.89	9
URL	0.81	0.55	0.65	31
VEHICLE_REGISTRATION_NUMBER	1.00	0.80	0.89	5
Accuracy	0.67 (775 samples)			
Macro Average	0.62	0.68	0.61	775
Weighted Average	0.70	0.67	0.66	775
Overall Accuracy	0.67			
Balanced Accuracy	0.73			
Macro F1-Score	0.61			
Micro F1-Score	0.67			
Matthews Correlation Coefficient	0.60			
Cohen's Kappa	0.60			

Table 13: RoBerTA Fine-Tuned Performance Metrics

Label	Precision	Recall	F1-Score	Support
ACTIVITIES	0.60	0.86	0.71	7
ADDRESS	0.78	0.74	0.76	19
BIRTH_DATE	1.00	1.00	1.00	6
COUNTRY	0.71	0.83	0.77	12
CRYPTO	0.75	0.94	0.83	16
CV_RESUME	1.00	0.94	0.97	16
DATE_TIME	0.67	0.70	0.68	20
EDUCATION_INFORMATION	0.91	1.00	0.95	10
EMAIL_ADDRESS	0.92	1.00	0.96	24
GEO_LOCATION	0.74	0.93	0.82	15
HEALTH_INSURANCE_ID	0.67	1.00	0.80	2
HOME_ADDRESS	0.73	0.73	0.73	15
JOB_TITLE	0.78	0.90	0.84	20
LANGUAGE	0.85	0.88	0.86	25
MEDICAL_HISTORY	1.00	0.95	0.98	21
NATIONALITY_CITIZENSHIP	1.00	1.00	1.00	3
NON_PI	0.88	0.84	0.86	326
NUMBER_OF_CHILDREN	0.75	1.00	0.86	3
ONLINE_IDENTIFIERS	0.80	0.80	0.80	10
PASSWORD	0.88	1.00	0.93	14
PERSON	0.89	0.89	0.89	9
PERSON_GENDER	1.00	1.00	1.00	2
PERSON_HEIGHT	1.00	1.00	1.00	4
PERSON_WEIGHT	0.80	1.00	0.89	4
PHONE_NUMBER	0.67	1.00	0.80	2
PICTURE_FACE	1.00	0.80	0.89	5
PLACE_OF_BIRTH	0.89	1.00	0.94	8
SCREEN_NAME	1.00	0.89	0.94	9
SHOPPING_BEHAVIOR	1.00	0.50	0.67	10
UNKWOWN	0.71	0.62	0.67	40
URL	0.96	0.87	0.92	31
VEHICLE_REGISTRATION_NUMBER	1.00	1.00	1.00	5
ZIPCODE	0.81	0.88	0.84	24
Accuracy	0.85 (737 samples)			
Macro Average	0.85	0.89	0.87	737
Weighted Average	0.86	0.85	0.85	737
Overall Accuracy	0.85			
Balanced Accuracy	0.89			
Macro F1-Score	0.87			
Micro F1-Score	0.85			
Matthews Correlation Coefficient	0.81			
Cohen's Kappa	0.81			