VASH: Evaluating Vagueness in Privacy Policies via Semantic Hierarchy Graph

Ziyan Zhou[1,2,3], Haoyang Yu[1,2,3], Jingjing Ma[1,2,3], Yanru He✉[1,3] ⋆, Yunchuan Guo[1,2,3], Liang Fang[1,3], Lingcui Zhang[1,3], and Fenghua Li[1,2,3]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] State Key Laboratory of Cyberspace Security Defense
{zhouziyan,heyanru,guoyunchuan}@iie.ac.cn

Abstract. Privacy policies serve as the primary mechanism for companies to disclose their data processing practices. However, vague language in these policies may expose users to risks. Existing methods for evaluating vagueness often struggle to capture fine-grained distinctions in vagueness among terms, lack sufficient consideration for domain-specific terminology, and neglect the phenomenon of false vagueness. To this end, we propose **VASH**, a **VA**gueness evaluation method for privacy policies based on the **S**emantic **H**ierarchy graph (SHGraph). We first extract terms from privacy policies across 475 industries to ensure comprehensive domain coverage. Leveraging these extracted terms and their hypernym–hyponym relationships, we construct a SHGraph to quantify the vagueness of individual terms based on their position within the semantic hierarchy. Subsequently, we identify false vagueness using the SHGraph and cue words, adjusting term vagueness scores accordingly. Finally, we apply attention weights to capture the influence of terms, thereby deriving sentence-level vagueness scores. To evaluate the effectiveness of VASH, we construct a large-scale dataset of 75,145 privacy policies. Extensive evaluations demonstrate that VASH effectively identifies both vagueness-related terms and false vagueness, yielding results highly consistent with human judgment. Moreover, through the application of VASH to our constructed dataset, we derive four findings that offer insights for enhancing privacy policy clarity. The source code and implementation details are available in our repository [4].

Keywords: Privacy policy · Vagueness evaluation · LLM.

1 Introduction

Privacy policies are legal documents that describe the data processing practices of applications, websites, and other services and aim to inform users about how their personal data are collected, used, and shared [7]. Data protection laws and regulations in various countries, such as the General Data Protection Regulation

⋆ Corresponding author: heyanru@iie.ac.cn

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

(GDPR) [2], explicitly require that privacy policies be clearly stated. However, to maintain long-term applicability of privacy policies in rapidly evolving technological and commercial environments, companies may use vague language to generalize their data practices [9]. More concerningly, some companies use vague terms as a dark pattern to obscure users' understanding of data processing and thereby seek greater control over personal information, which in turn erodes user trust and damages companies' reputations [11]. Therefore, it is crucial to effectively evaluate the vagueness of privacy policies.

Existing methods for privacy policy vagueness evaluation fall into two categories: Neural Network-based methods [16, 10] and term-based methods [9, 11, 20, 19, 17]. Neural Network-based methods aim to directly predict vagueness scores, but their limited interpretability restricts their practical applicability. Moreover, the performance of these models is constrained by the scarcity of high-quality annotated data and a tendency to capture superficial statistical patterns rather than achieving genuine semantic comprehension. In contrast, term-based methods enhance interpretability by explicitly identifying vague terms. However, they quantify vagueness by only counting vague terms, which can lead to inaccurate assessments. Such methods also fail to account for cases in which vagueness is mitigated through the provision of specific examples or clarifying details. Moreover, their inability to effectively recognize domain-specific terminology further limits the accuracy of vague term identification and overall assessment.

To address the above problems, we propose VASH, a fine-grained vagueness evaluation method for privacy policies. Our main contributions are as follows:

- We propose VASH, a vagueness evaluation method for privacy policies that integrates semantic hierarchy analysis with false vagueness detection, enabling more accurate assessments.
- To enable fine-grained assessment, we design a semantic hierarchy graph that captures distinctions in term vagueness through hypernym-hyponym relationships and incorporates attention weights to measure terms' impact on sentence vagueness. To further improve accuracy, we leverage SHGraph and cue words to calibrate vagueness scores under contextual clarification.
- We construct a large-scale dataset of 75,145 privacy policies across 475 industries from 2016 to 2025. Extensive experiments on this dataset demonstrate VASH's effectiveness, achieving an F1-score of 0.82 for false vagueness detection, while identifying an average of 100.21 vagueness-related terms and 7.84 instances of false vagueness per policy. Moreover, our analysis reveals four findings that offer valuable insights into privacy policy clarity.

2 Related Work

2.1 Graph-Based Analysis of Privacy Policies

Graph-based representations, including ontologies and knowledge graphs, are increasingly used to analyze privacy policies. Specifically, ontologies represent semantic relationships among privacy terms [3, 8], while knowledge graphs model

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

interactions between data types and entities [12]. Regarding construction methodologies, prior work typically employs neural networks for relation prediction [14] or rule-based syntactic patterns for hypernym extraction [13]. Although these structures facilitate downstream tasks such as vague expression identification [17], existing privacy ontologies suffer from limited scale and coverage, resulting in imprecise analyses. To bridge this gap, we extract terms from a large-scale privacy policy dataset to construct a semantic hierarchy graph comprising 8,229 terms, providing a robust foundation for accurate vagueness analysis.

2.2 Vagueness Evaluation of Privacy Policies

As shown in Section 1, existing approaches can be categorized into Neural Network-based methods [16, 10] and term-based methods [9, 11, 20, 19, 17]. Neural Network-based methods typically train classifiers on annotated datasets to predict sentence-level vagueness, aggregating these predictions into an overall score. However, these methods face two limitations: (i) manual annotation suffers from poor inter-rater consistency [16]; (ii) the black-box nature of deep neural networks obscures the rationale behind predictions. In contrast, term-based methods employ NLP techniques to identify vague terms and quantify vagueness via the proportion of vague terms or sentences. While this enhances interpretability, most studies assume uniform vagueness across terms. Andow et al. [9] attempt to refine this by organizing vague terms in a tree-structured ontology, where vagueness depends on distance to the root, but they do not capture complex semantic relationships such as multiple inheritance. Moreover, term-based methods struggle with *false vagueness*. Lian et al. [18] introduce a rule-based method to detect false vagueness, but their approach lacks generalizability and does not integrate false vagueness detection into the vagueness quantification metric. To overcome these challenges, we construct a semantic hierarchy graph featuring extensive term coverage and rich semantic relations to differentiate varying levels of vagueness among terms. By utilizing this structure to identify false vagueness and dynamically adjust term contributions based on explanatory relationships, our approach achieves a more precise evaluation.

3 Methodology

3.1 Overview

As shown in Figure 1, VASH comprises three modules. The **dataset construction module** constructs two datasets, TermEx and PriPolicy, by collecting privacy policies together with their historical versions. The **semantic hierarchy graph generation module** extracts vagueness-related terms from TermEx, uses LLMs to identify their semantic relations, and organizes them into the SHGraph where each node represents a term annotated with a vagueness score. The **vagueness scoring module** generates a vagueness score for each policy in PriPolicy through a multi-step process: the SHGraph is applied to detect false vagueness, sentence-level vagueness is computed using attention weights, and sentence-level scores are averaged across the policy.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

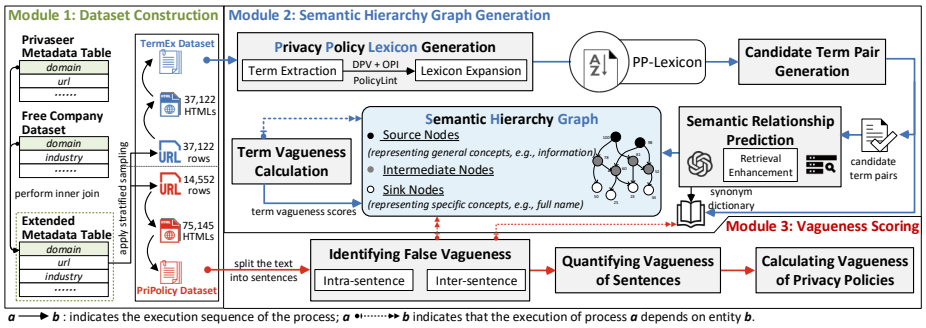


Fig. 1: Overview of VASH.

3.2 Dataset Construction

To facilitate the construction of SHGraph and the application of VASH for analyzing temporal and cross-industry variations in privacy policy vagueness, we compile two datasets: TermEx and PriPolicy. TermEx dataset contains 37,122 latest-version policies across 475 industries for term extraction, while PriPolicy dataset comprises 75,145 policies from the same industries spanning 2016 to 2025 for evaluation. These collections address the limitations of existing datasets [21, 16, 19, 11, 9, 7], specifically their lack of industry diversity and post-2020 coverage. By incorporating comprehensive, up-to-date policies, our datasets capture the impact of regulatory changes on policy clarity and reveal how companies adjust their disclosures. The construction process involves four steps: (1) merging the metadata from Privaseer [5] and the Free Company Dataset [1] based on the shared *domain* field to create an extended metadata table; (2) applying stratified sampling by industry to select privacy policy *urls* for PriPolicy (14,552 entries) and TermEx (37,122 entries); (3) retrieving historical snapshots via the Wayback Machine [6]; and (4) extracting the textual content from each snapshot.

3.3 Semantic Hierarchy Graph Generation

SHGraph assigns vagueness scores to terms to facilitate the evaluation of privacy policies. To construct the SHGraph, we first extract terms from the TermEx to build the PP-Lexicon, which constitutes the nodes of the SHGraph. Then we identify candidate pairs exceeding a semantic similarity threshold and utilize an LLM, enriched with context chunks from PriPolicy, to determine their semantic relations. Finally, we compute a vagueness score for each term based on its structural position within the SHGraph.

Privacy Policy Lexicon Generation To obtain PP-Lexicon, we first extract candidate terms using 30 part-of-speech patterns[1] that capture specific linguistic

[1] The full list of patterns is available in our repository.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

structures (e.g., "NN NN" for two consecutive nouns such as "email address"). The candidates then undergo a filtering process involving spelling correction, normalization, and removal of rare or redundant terms, reducing the lexicon to approximately 20,000 terms for manual review. To ensure comprehensive coverage, we further augment this list by aligning it with existing privacy ontologies [15, 8, 3]. The final curated PP-Lexicon comprises 8,229 terms tailored for vagueness analysis.

Candidate Term Pair Generation To mitigate the computational inefficiency of directly predicting semantic relationships for all possible term combinations in the PP-Lexicon, we implement a two-step strategy to obtain candidate pairs with a high probability of exhibiting semantic relationships:

- *Synonym Replacement.* To reduce redundancy, we perform synonym replacement by calculating cosine similarities between term embeddings. Pairs exceeding a threshold of 0.9 are manually examined to verify synonymy and clustered into disjoint sets. Within each set, the term with the highest cumulative similarity to all others serves as the representative central term.
- *Candidate Selection via Transitive Closure.* Then, we leverage the transitive closure property to iteratively select pairs that may exhibit semantic relationships. Formally, let $\mathcal{T}$ denote the PP-Lexicon. We define a direct relation $D \subseteq \mathcal{T} \times \mathcal{T}$ as $D = \{(t_i, t_j) \in \mathcal{T} \times \mathcal{T} \mid \cos\_\text{sim}(t_i, t_j) \geq \theta\}$, where θ is empirically set to 0.7. For a term $t \in \mathcal{T}$, we define its associated candidate set R_t through an iterative process: (i) initialize $R_t^{(0)} = \{s \in \mathcal{T} \mid (t, s) \in D\}$; and (ii) for $k \geq 0$, update the set as $R_t^{(k+1)} = R_t^{(k)} \cup \{s \in \mathcal{T} \mid \exists r \in R_t^{(k)} : (r, s) \in D\}$. The final candidate set for term t is $R_t = \bigcup_{k=0}^{\infty} R_t^{(k)}$. As $\mathcal{T}$ is finite, this process converges after a finite number of steps. Finally, we construct the comprehensive collection of potential pairs $C = \bigcup_{t \in \mathcal{T}} \{< x, y > \mid x, y \in R_t, x \neq y\}$. This ensures that C encompasses all pairs likely to exhibit semantic relationships, thereby substantially reducing the total number of pairs requiring prediction.

Semantic Relationship Prediction We leverage GPT-4o to predict semantic relationships between terms in each candidate pair. To enhance prediction accuracy, we implement a structured prompting approach that incorporates Retrieval-Augmented Generation (RAG). As shown in Figure 2, the prompt comprises three components: Context, Task, and Action, which guide the model from definitional comprehension to relationship inference. External knowledge is integrated through a vector database containing term definitions extracted from privacy policies. For each term in a given pair, the most relevant chunk is retrieved and incorporated into the prompt. Formally, let $\mathcal{R} : \mathcal{T} \times \mathcal{T} \to \{-1, 0, 1, 2\}$ denote the relationship function predicted by the LLMs, where $\mathcal{T}$ is the PP-Lexicon. For any given pair (t_1, t_2), $\mathcal{R}(t_1, t_2) = 1$ indicates that t_1 is a hypernym of t_2, -1 a hyponym, 0 a synonym, and 2 the absence of a semantic relation. Based on these predictions, we construct two knowledge structures: (i) a synonym dictionary comprising 1,243 terms (including 467 central terms), and (ii) a directed acyclic graph (DAG) consisting of 7,453 nodes and 11,318 edges.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

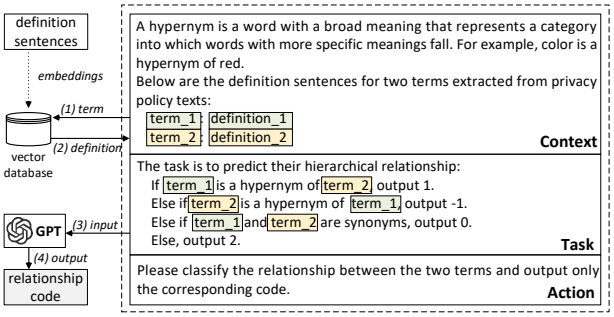


Fig. 2: Semantic relationship prediction.

Term Vagueness Calculation We construct the SHGraph to measure semantic distances between terms and quantify term vagueness in privacy policies. To formalize our methodology, we first define the *SHGraph* and its *reachable source/sink nodes*. Subsequently, building upon this graph structure, we introduce the concept of *semantic chains*. This concept captures the semantic flow from abstract to concrete terms and aligns with the intuition that general concepts typically exhibit greater vagueness than their specific descendants.

Definition 1 (Semantic Hierarchy Graph). *The semantic hierarchy graph is formalized as a directed acyclic graph, denoted as* $SHGraph = (V, E, f)$*. Here,* V *is the set of nodes corresponding one-to-one with terms in the PP-Lexicon* $\mathcal{T}$*,* E *represents the set of directed edges, and* $f : V \to [0, 100]$ *assigns a vagueness score to each node, with higher values indicating greater vagueness. The nodes in* V *are categorized into three subsets: (i) Source Nodes* S*, where each node* $v \in S$ *has no incoming edges but at least one outgoing edge, representing a general concept (e.g., information); (ii) Sink Nodes* T*, where each node* $v \in T$ *has no outgoing edges but at least one incoming edge, representing a specific concept (e.g., last name); and (iii) Intermediate Nodes* I*, where each node* $v \in I$ *has both incoming and outgoing edges.*

Definition 2 (Reachable Source/Sink Node). *Given* $SHGraph = (V, E, f)$ *and a node* $v \in V$*, a source node* $v_s \in S$ *is a reachable source node of* v *if there exists a directed path from* v_s *to* v*; similarly, a sink node* $v_t \in T$ *is a reachable sink node of* v *if there exists a directed path from* v *to* v_t*. We denote the sets of reachable source and sink nodes of* v *by* $RS(v)$ *and* $RT(v)$*, respectively.*

Definition 3 (Semantic Chain). *Given* $SHGraph = (V, E, f)$*, a semantic chain is a directed path* $p = (v_1, v_2, \ldots, v_k)$ *if (i)* $v_1 \in S$ *is a source node and* $v_k \in T$ *is a sink node; (ii)* $(v_i, v_{i+1}) \in E$ *for all* $1 \leq i < k$*; and (iii) for any consecutive nodes* v_i *and* v_{i+1} *in the path,* $f(v_i) \geq f(v_{i+1})$*. For example, a semantic chain might follow the path: information* $\to \ldots \to$ *email address, where each successive term is more specific and less vague than its predecessor.*

Leveraging the established hierarchical structure, we quantify the vagueness of each term. Intuitively, a node with numerous reachable source nodes inherits

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

semantics from multiple superordinate categories, resulting in greater semantic constraint and reduced vagueness. Conversely, a high number of reachable sink nodes implies that the node generalizes over a broader scope, corresponding to increased vagueness. Moreover, the longest path distance to sink/source nodes further reflects how far a concept extends across semantic levels, thereby capturing the hierarchical differentiation of vagueness. Therefore, the vagueness of a node can be inferred from how broadly and how deeply it connects to other nodes along the semantic hierarchy. As shown in Eq. (1), we compute the vagueness score $f(v)$ of a node v based on the topological properties of the SHGraph.

$$f(v) = \alpha \cdot f_s(v) + (1 - \alpha) \cdot f_t(v), \tag{1}$$

where $f_s(v)$ and $f_t(v)$ quantify the influence of reachable source nodes and sink nodes on the vagueness of node v, respectively, as defined in Eqs. (2) and (3).

$$f_s(v) = \begin{cases} \frac{norm_s}{\sqrt{sv \times sp}} \times 100 & \sqrt{sv \times sp} \geq norm_s, \\ 100 & \sqrt{sv \times sp} < norm_s \end{cases} \tag{2}$$

$$f_t(v) = \begin{cases} \frac{\sqrt{tv \times tp}}{norm_t} \times 100 & \sqrt{tv \times tp} \leq norm_t, \\ 100 & \sqrt{tv \times tp} > norm_t \end{cases} \tag{3}$$

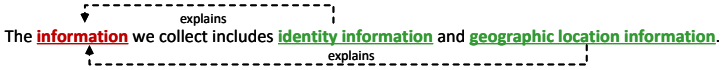
where $sv = |RS(v)|$ and $tv = |RT(v)|$ denote the numbers of reachable source and sink nodes of v, respectively. The parameters $sp = \max_{x \in RS(v)} Len(v, x)$ and $tp = \max_{x \in RT(v)} Len(v, x)$ are the longest distances from v to the nodes in $RS(v)$ and $RT(v)$, respectively. To mitigate the impact of extreme values, we implement truncation-based normalization functions, $f_s(v)$ and $f_t(v)$, which employ thresholds $norm_s$ and $norm_t$ to cap extreme values. Moreover, we apply a square-root transformation to $\sqrt{sv \times sp}$ and $\sqrt{tv \times tp}$; this reduces the marginal impact of outliers while preserving the nonlinear growth of semantic generality across hierarchical levels. This design ensures that vagueness scores transition smoothly and meaningfully along the semantic hierarchy. Furthermore, the weighting parameter $\alpha = \frac{sv'}{sv' + tv'}$ in Eq. (1) dynamically balances upper and lower semantic connections using clipped values sv' and tv', which are capped by truncation thresholds ϕ_s and ϕ_t, respectively. Specifically, if $sv > \phi_s$, then sv' is set to ϕ_s; similarly, if $tv > \phi_t$, then tv' is set to ϕ_t.

To determine the optimal values for the four parameters ($norm_s$, $norm_t$, ϕ_s, and ϕ_t) in Eq. (1), we analyze the distributions of structural indicators (sv_i, tv_i, $\sqrt{sv_i \times sp_i}$, and $\sqrt{tv_i \times tp_i}$) across the SHGraph. A heuristic search is then conducted to identify parameter sets that ensure monotonic vagueness decreases along semantic chains and minimize the coefficient of variation.

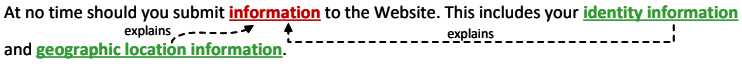
3.4 Vagueness Scoring

Identifying False Vagueness When a general term (hypernym) is explained by a more specific term (hyponym), we refer to this phenomenon as false vague-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15



(a) False intra-sentence vagueness.



(b) False inter-sentence vagueness.

Fig. 3: False vagueness in sentences.

ness. False vagueness does not imply the absence of vagueness; instead, it describes a situation in which the model overestimates the actual degree of vagueness. For clarity, we denote the general term as the *explained term* and the term providing the interpretation as the *explaining term*. Sentences containing *explained terms* are referred to as *explained sentences*. To identify false vagueness, we first mark candidate sentences using cue words (e.g., *"such as"*), followed by a part-of-speech analysis to extract terms matching the 30 part-of-speech patterns (Section 3.3). Then, we use the SHGraph to validate semantic relationships between extracted terms, addressing both intra-sentence and inter-sentence contexts (Figure 3). Upon validation, the vagueness scores of the *explained terms* are adjusted downward to reflect their true vagueness within that sentence.

Quantifying Vagueness of Sentences We categorize sentences into standard sentences (S_{non}) and those exhibiting false vagueness (S_{fv}). Let $\mathcal{T}_s \subseteq \mathcal{T}$ denote terms extracted from sentence s. For a sentence $s \in S_{non}$, its vagueness score is determined by the vagueness score of each term $t \in \mathcal{T}_s$ derived from the SHGraph, and the contextual importance of each term within the sentence. We quantify term importance using attention weights from BERT's last layer, which dynamically assign scores to tokens according to their contextual contribution. Formally, the vagueness score of sentence s is defined as follows:

$$Vag(s) = \frac{1}{\Omega}\left(\sum_{t\in\mathcal{T}_s} \omega_t \cdot f(t)\right), \tag{4}$$

where $f(t)$ is the vagueness score of term t, ω_t is the attention weight of term t, and $\Omega = \sum_{t\in\mathcal{T}_s} \omega_t$ is the sum of the attention weights of all identified terms.

For a sentence $s \in S_{fv}$, we update the vagueness scores of the *explained terms* in the sentence to enable a more accurate evaluation. For each *explained term* t in sentence s, the updated vagueness score of t is defined as the average of its original score $f(t)$ and the scores of its *explaining terms*, as shown in Eq. (5):

$$Vag(t) = \frac{f(t) + \sum_{t'\in\mathcal{T}_t} f(t')}{|\mathcal{T}_t| + 1}, \tag{5}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

where $\mathcal{T}_t$ is the set of *explaining terms* for t, and $t' \in \mathcal{T}_t$ refers to an *explaining term* of t. The vagueness score of the sentence s is then computed using Eq. (6):

$$Vag(s) = \frac{1}{\Omega}\left(\sum_{t_i \in \mathcal{T}_s^{ex}} \omega_{t_i} \cdot Vag(t_i) + \sum_{t_j \in \mathcal{T}_s \setminus \mathcal{T}_s^{ex}} \omega_{t_j} \cdot f(t_j)\right), \tag{6}$$

where the updated scores $Vag(t)$ are applied to *explained terms* and the original scores $f(t)$ are used for all other terms. Here, $\mathcal{T}_s^{ex} \subseteq \mathcal{T}_s$ denotes the set of *explained terms* in s, and $\Omega = \sum_{t \in \mathcal{T}_s} \omega_t$ is the total attention weight of terms in s.

Calculating Vagueness of Privacy Policies Given a privacy policy p, its overall vagueness score $Vag(p)$ is computed by averaging the vagueness scores of all its sentences: $Vag(p) = \frac{\sum_{s \in S_p} Vag(s)}{|S_p|}$, where S_p represents the set of sentences in policy p with vagueness scores greater than zero, and $Vag(s)$ denotes the vagueness score of the sentence $s \in S_p$.

4 Experimental Evaluation

We conduct a series of experiments to evaluate the effectiveness of VASH. Our evaluation consists of two parts: **(i) comparative experiments,** which measure VASH's accuracy in semantic relationship prediction (Q1), vagueness-related term identification (Q2), and false vagueness detection (Q3) against state-of-the-art baselines. **(ii) an effectiveness evaluation**, which investigates the overall performance of VASH through a survey-based study.

4.1 Comparative Experiments

Datasets. We use three datasets for evaluations. First, for semantic relationship prediction, we curate a test set of 1,000 term pairs randomly sampled from the candidate pairs in Section 3.3. Second, we leverage the PriPolicy dataset to assess SHGraph's capability in identifying vagueness-related terms. Finally, for false vagueness detection, we construct a dataset of 200 sentences from PriPolicy, half of which are flagged by SHGraph, OPI [3], or PolicyLint [8]. To ensure ground truth quality, both the first and third datasets are subjected to independent review by three annotators, with disagreements resolved via majority voting.
Metrics. For semantic relationship prediction, we employ standard classification metrics including Precision, Recall, and F1-score. For vagueness-related term identification, we report the average number of detected vagueness-related terms per policy ($term_{avg}$) as an indicator of detection sensitivity. For false vagueness detection, we evaluate performance using Precision, Recall, and F1-score, as well as the average number of detected *explained sentences* per policy ($exsent_{avg}$).
Baselines. Since no existing method fully aligns with our objectives, we compare VASH against approaches with partially related goals. First, for semantic

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

Table 1: Comparison of tools in identifying vagueness-related terms and detecting `explained sentences`.

Tool	$term_{avg}$	$exsent_{avg}$	Precision	Recall	F1-Score
DPV	39.66	-	-	-	-
PolicyLint	24.17	3.88	0.69	0.58	0.63
OPI	**111.92**	4.30	0.72	0.37	0.49
SHGraph	100.21	**7.84**	**0.74**	**0.92**	**0.82**

relationship prediction, we compare our method with a state-of-the-art Convolutional Neural Network (CNN)-based method [14] specifically designed for privacy policy analysis. For vagueness-related term identification, we compare SHGraph with three widely-used privacy policy analysis tools: DPV [15], PolicyLint [8], and OPI [3]. For false vagueness detection, we compare our approach against PolicyLint, OPI, and the method proposed by Lian et al. [18].

Results and Analysis. ▶ *Answer to Q1: Semantic Relationship Prediction.* Our prediction method demonstrates superior performance across all metrics, achieving an overall F1-score of 0.80, indicating the effectiveness of leveraging advanced LLMs to predict semantic relations between terms. Specifically, GPT-4o maintains a balanced performance with high Precision (> 0.75) and Recall (> 0.77) for both hypernym-hyponym and unrelated pairs. In contrast, the baseline struggles to distinguish subtle semantic boundaries, resulting in a lower Recall (0.47) for hypernym-hyponym relation identification.

▶ *Answer to Q2: Vagueness-Related Term Identification.* Table 1 shows that SHGraph identifies 100.21 vagueness-related terms per policy on average, outperforming DPV (39.66) and PolicyLint (24.17), while approaching OPI (111.92). Although SHGraph detects slightly fewer terms than OPI, it focuses on terms that are truly vagueness-related, whereas OPI may include broader privacy terms that are not necessarily related to vagueness. For example, OPI identifies *"privacy policy"* as a vagueness-related term, but this is standard terminology in privacy policies, and including it may impact vagueness evaluation accuracy.

▶ *Answer to Q3: False Vagueness Detection.* As shown in Table 1, SHGraph demonstrates superior performance in detecting false vagueness. On average, it identifies 7.84 explained sentences per policy, approximately doubling the detection rates of PolicyLint (3.88) and OPI (4.30). When comparing with the approach proposed by Lian et al. [18] , which reported precision, recall, and F1-scores of 0.71, 0.68, and 0.69 respectively, our method demonstrates consistent superiority across all metrics (0.74, 0.92, and 0.82). Moreover, rather than merely detecting false vagueness, VASH integrates these findings to recalibrate the overall vagueness assessment, thereby resulting in more reliable evaluations.

4.2 Survey-based Evaluation

To validate the alignment of our vagueness assessment with human judgment, we conduct a dual evaluation consisting of a user study and an LLM-based evaluation. Specifically, the evaluation focuses on answering two main questions:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

(Q4) Can VASH identify vagueness-related terms in privacy policies? (Q5) Are the vagueness evaluations produced by VASH consistent with human judgment?

The evaluation involves 44 human participants and three LLMs (GPT-4o, GPT-4-turbo, and Claude-3.5-Sonnet). For convenience, we collectively refer to both human participants and LLMs as *participants*. All participants complete an identical questionnaire comprising two tasks with 10 questions each, numbered T1.1 to T1.10 and T2.1 to T2.10. Notably, each question in Task 1 contains two components, denoted as sub-questions a and b (e.g., T1.1(a) and T1.1(b)). The specific descriptions of the tasks are as follows:

- **Task 1 (Term Identification):** Participants are presented with ten sentences from real-world privacy policies, in which terms identified by VASH are highlighted. For each sentence, participants answer two sub-questions: (a) *Did the highlighted terms influence your judgment of the sentence's vagueness?* (`Yes`/ `No`/ `Uncertain`); and (b) *How vague do you consider the sentence?* (Rated on a five-point Likert scale, where higher scores indicate greater vagueness).
- **Task 2 (Comparative Judgment):** Participants review ten pairs of comparable sentences and are asked to identify the vaguer option (Sentence A or B), or select `Uncertain` if they are unable to decide.

Results. We collect responses from both the human participants (mean completion time: 838 seconds) and the three LLMs. To investigate discrepancies where participant responses deviated from VASH evaluations, we implement a follow-up mechanism: conducting interviews with human participants and applying an analogous querying protocol to the LLMs. The combined findings from the questionnaires and these supplementary investigations are detailed below.

▶ *Results of Task 1 (Term Identification).* Results from sub-question (a) indicate that 86.36% of human participants agree with VASH's identification of vagueness-related terms, which is further confirmed by the LLMs. For sub-question (b), we compute the average ratings for human and LLM participants separately. To facilitate comparison, we normalize VASH's 0–100 scores to a 0–5 scale by dividing by 20, aligning them with the 5-point Likert scale. As shown in Figure 4, the overall trend of the VASH scores is broadly consistent with the ratings given by humans and LLMs. One notable exception occurs in Sentence 8, *"we will not sell or rent your personal data to third parties for marketing purposes."* While participants rate this statement as clear or neutral, VASH classifies it as highly vague. Legal expert review confirms VASH's accuracy due to the lack of specific definitions for key terms like *"personal data"* and *"marketing purposes"* .

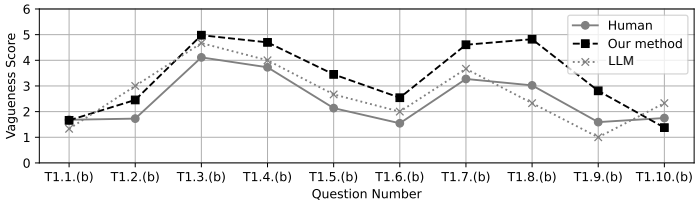


Fig. 4: Participant average ratings vs. vagueness scores computed by VASH.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

No.	Sentence A	Sentence B
5	Browsing information: Information related to your interactions with our Websites, including IP addresses and browser behavior, to analyze trends, administer the Website, track user activities and movements across the Website, infer user interests, and otherwise induce, deduce, and gather information about individual users and market segments.	Browsing information: Information related to your interactions with our Websites, our marketing emails or other communications, including information such as IP addresses and browser behavior, to analyze trends, administer the Website, track user activities and movements across the Website, infer user interests, and otherwise induce, deduce, and gather information about individual users and market segments.
7	We may collect basic user profile information from all of our Visitors.	We may also collect information on how the Service is accessed and used ("Usage Data").

Fig. 5: Sentence pairs of T2.5 and T2.7.

This demonstrates that VASH possesses superior sensitivity to latent vagueness, effectively identifying subtle yet critical vague elements in privacy policies.

▶ *Results of Task 2 (Comparative Judgment).* As shown in Table 2, VASH demonstrates strong alignment with the majority of human and LLM participants, with primary divergences observed in T2.5 and T2.7 (Figure 5). In T2.5, 65.91% of human participants judge Sentence A as more vague, consistent with VASH. Conversely, GPT-4o and Claude-3.5-Sonnet argue that Sentence B is vaguer due to its broader data collection scope, while GPT-4-turbo remain undecided. Regarding T2.7, human preferences for A and B are nearly distinctively balanced, and Claude-3.5-Sonnet selected `Uncertain`. In fact, the vagueness scores computed by VASH are also very close, with sentence A receiving a score of 93.91 and sentence B 99.63, demonstrating that VASH can effectively capture subtle differences in vagueness and reflect inherent uncertainty.

5 Findings

We apply VASH to the PriPolicy dataset to examine vagueness across three dimensions: time, company size, and industry. The following findings characterize the specific impact of these factors on privacy policy vagueness.

Finding 1 (Temporal Perspective): Over the past decade, regulatory pressure has driven increases in policy length and explanatory content, but overall clarity has barely improved. Table 3 shows that while the average policy length (60.06 to 67.42) and the number of `explained sentences` (6.74 to 7.69) increase between 2016 and 2025, the average vagueness score improves by only 0.14 points. Notably, 2019 exhibits significant changes in policy modifications. The proportion of companies adding `explained sentences` is twice that of 2017 and 1.24 times that of 2018. Moreover, the proportion of

Table 2: LLM responses for Task 2. 'A' and 'B' indicate that Sentence A or B is more vague, respectively, while 'U' denotes uncertainty.

Participant/Method	T2.1	T2.2	T2.3	T2.4	T2.5	T2.6	T2.7	T2.8	T2.9	T2.10
GPT-4o	A	A	A	A	B	A	A	B	B	B
GPT-4-turbo	A	B	A	A	U	A	A	B	B	B
Claude-3.5-Sonnet	A	B	A	A	B	A	U	A	B	B
Human participants	**77.27% (A)** 18.18% (B) 4.55% (U)	**61.36% (A)** 31.82% (B) 6.82% (U)	**61.36% (A)** 31.82% (B) 6.82% (U)	**81.82% (A)** 6.82% (B) 11.36% (U)	**65.91% (A)** 25.00% (B) 9.09% (U)	**79.55% (A)** 13.64% (B) 6.82% (U)	**43.18% (A)** **36.36**% (B) 20.45% (U)	15.91% (A) **75.00% (B)** 9.09% (U)	15.91% (A) **68.18% (B)** 15.91% (U)	9.09% (A) **84.09% (B)** 6.82% (U)
Ours	A	A	A	A	A	A	B	B	B	B

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

Table 3: Temporal evolution of privacy policy metrics and their changes after updates (2016–2025). Here, $vscore_{avg}$ and $vscore_{var}$ denote the average and variance of vagueness scores, respectively, and $exsent_{avg}$ denotes the average number of `explained sentences`. For policy updates, $\%\_update$ shows the proportion of policies updated, while $\%\_vscore_{dec}$, $\%\_exsent_{inc}$, and $\%\_len_{inc}$ indicate the proportion of updates that resulted in decreased vagueness scores, increased `explained sentences`, and increased policy length, respectively.

Year	$vscore_{avg}$	$vscore_{var}$	$exsent_{avg}$	len_{avg}	$\%\_update$	$\%\_vscore_{dec}$	$\%\_exsent_{inc}$	$\%\_len_{inc}$
2016	82.13	35.45	6.74	60.06	-	-	-	-
2017	82.13	35.35	6.76	60.23	2.41	51.00	23.06	42.17
2018	82.14	35.10	6.82	60.68	3.26	50.42	37.11	53.38
2019	82.13	35.04	6.98	62.02	5.88	50.76	**46.19**	**60.94**
2020	82.13	34.36	7.14	63.37	9.11	49.13	34.81	50.57
2021	82.13	33.88	7.24	64.09	9.45	50.55	29.13	44.36
2022	82.16	33.05	7.39	65.24	13.19	50.05	26.19	42.34
2023	82.16	33.40	7.51	66.11	12.73	48.00	25.82	41.20
2024	82.12	34.29	7.65	67.07	13.35	50.39	27.07	44.42
2025	81.99	39.27	7.69	67.42	18.49	51.30	28.47	42.34

companies extending their privacy policies is 1.45 times that of 2017 and 1.14 times that of 2018. This substantial increase is likely driven by the enforcement of the GDPR in 2018. Furthermore, the decline in update metrics ($\%\_len_{inc}$ and $\%\_exsent_{inc}$) after 2020 further indicates a lack of sustained motivation to optimize privacy policies during periods of reduced regulatory oversight.

Finding 2 (Company Size Perspective): Company size does not significantly affect privacy policy vagueness, but it does lead to an increase in explanatory content. Table 4 shows no significant variation in privacy policy vagueness across company sizes. This finding is supported by a one-way ANOVA conducted on 2025 data (F=1.17, p=0.32). We attribute this uniformity to counterbalancing factors. Large enterprises (LEs) face greater operational complexity, which may increase the use of vague language, whereas small and medium-sized enterprises (SMEs) benefit from narrower scopes that facilitate clearer expression despite more limited compliance resources. In contrast, company size is linked to differences in explanatory depth. LEs include more `explained sentences` on average, indicating a greater tendency to provide clarifications that mitigate potential vagueness.

Finding 3 (Industry Category Perspective): Privacy policy vagueness and explanatory practices have significant influenced by industry category. A one-way ANOVA indicates significant differences in policy vagueness across industries (F=4.64, p=7.63×10^{-11}). The *Oil, Gas, and Mining* sector exhibits the highest vagueness scores and the lowest variance, whereas *Hospitals*

Table 4: Comparison of policy vagueness by size (2025).

Size	$vscore_{avg}$	$vscore_{var}$	$exsent_{avg}$	len_{avg}
SMEs (1–500)	81.98	38.91	7.63	66.89
LEs (500+)	82.14	46.07	9.01	77.50

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

and Health Care shows the lowest vagueness, likely reflecting strict regulations on sensitive health data. In terms of policy length and explanatory content, *Technology, Information, and Media* leads with an average of 79.40 sentences and 10.20 `explained sentences` per policy, while traditional industries rank lowest on both measures. These patterns may reflect differential regulatory scrutiny, with technology-intensive sectors under stronger pressure to clarify data usage compared to traditional industries.

Finding 4 (Quality Over Quantity): Privacy policy updates do not guarantee clarity improvements, and the degree of enhancement depends on the quality of updated content. Table 3 shows that only $\approx 50\%$ of updates result in reduced vagueness scores. Moreover, outside the 2018–2020 period, fewer than 30% of updates increase `explained sentences`, and only $\approx 43\%$ extend policy length. These findings suggest that nearly half of companies make only superficial adjustments to their privacy policies to satisfy minimal regulatory requirements, rather than enhancing policy clarity.

6 Conclusion

In this paper, we propose VASH, a novel method that employs the semantic hierarchy graph (SHGraph) to quantify vagueness in privacy policies. By extracting and organizing terms into hierarchical structures, SHGraph captures fine-grained vagueness distinctions among terms. The integration of predefined cue words enables the identification of false vagueness and subsequent score recalibration, ensuring assessment accuracy. Evaluation on 75,145 privacy policies demonstrates VASH's effectiveness, identifying an average of 100.21 vagueness-related terms per policy and achieving an F1-score of 0.82 for false vagueness detection. The resulting vagueness scores also show high consistency with human judgment. Moreover, our analysis reveals regulatory pressure and industry norms drive policy lengthening but fail to consistently improve clarity. Future work will extend these findings by developing automated approaches for updating privacy policies to improve their transparency and comprehensibility for users.

7 Acknowledgements

This work was supported by the National Key R&D Program of China (No.2023YFB3106505).

References

1. Bigpicture 2023 q4 free company dataset. `https://www.kaggle.com/datasets/mfrye0/bigpicture-company-dataset/data`
2. General data protection regulation (gdpr). `https://gdpr-info.eu/`
3. Ontology of personal information. `https://opi.cs.cmu.edu/about.html\#`
4. Our anonymous repository. `https://anonymous.4open.science/r/vash-iccs`

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15

5. Privaseer corpus metadata. https://git.psu.edu/hlt-lab/PrivaSeer-Corpus/-/tree/main/metadata?ref_type=heads
6. Wayback machine. https://web.archive.org/
7. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., Mayer, J.: Privacy policies over time: Curation and analysis of a million-document dataset. In: Proceedings of the Web Conference 2021. pp. 2165–2176 (2021)
8. Andow, B., Mahmud, S.Y., Wang, W., Whitaker, J., Enck, W., Reaves, B., Singh, K., Xie, T.: {PolicyLint}: investigating internal privacy policy contradictions on google play. In: 28th USENIX security symposium (USENIX security 19). pp. 585–602 (2019)
9. Andow, B., Mahmud, S.Y., Whitaker, J., Enck, W., Reaves, B., Singh, K., Egelman, S.: Actions speak louder than words:{Entity-Sensitive} privacy policy and data flow analysis with {PoliCheck}. In: 29th USENIX Security Symposium (USENIX Security 20). pp. 985–1002 (2020)
10. Belcheva, V., Ermakova, T., Fabian, B.: Understanding website privacy policies—a longitudinal analysis using natural language processing. Information **14**(11), 622 (2023)
11. Bhatia, J., Breaux, T.D., Reidenberg, J.R., Norton, T.B.: A theory of vagueness and privacy risk perception. In: 2016 IEEE 24th International Requirements Engineering Conference (RE). pp. 26–35. IEEE (2016)
12. Cui, H., Trimananda, R., Markopoulou, A., Jordan, S.: {PoliGraph}: Automated privacy policy analysis using knowledge graphs. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 1037–1054 (2023)
13. Evans, M.C., Bhatia, J., Wadkar, S., Breaux, T.D.: An evaluation of constituency-based hyponymy extraction from privacy policies. In: 2017 IEEE 25th International Requirements Engineering Conference (RE). pp. 312–321. IEEE (2017)
14. Hosseini, M.B., Heaps, J., Slavin, R., Niu, J., Breaux, T.: Ambiguity and generality in natural language privacy policies. In: 2021 IEEE 29th International Requirements Engineering Conference (RE). pp. 70–81. IEEE (2021)
15. J. Pandit, H., Esteves, B., P. Krog, G., Ryan, P., Golpayegani, D., Flake, J.: Data privacy vocabulary (dpv)–version 2.0. In: International Semantic Web Conference. pp. 171–193. Springer (2024)
16. Lebanoff, L., Liu, F.: Automatic detection of vague words and sentences in privacy policies. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3508–3517 (2018)
17. Li, S., Yang, Z., Nan, Y., Yu, S., Zhu, Q., Yang, M.: Are we getting well-informed? an in-depth study of runtime privacy notice practice in mobile apps. In: Proceedings of the 31st ACM Conference on Computer and Communications Security (CCS 2024). pp. 1581–1595 (2024)
18. Lian, X., Huang, D., Li, X., Zhao, Z., Fan, Z., Li, M.: Really vague? automatically identify the potential false vagueness within the context of documents. Mathematics **11**(10), 2334 (2023)
19. Liu, F., Fella, N.L., Liao, K.: Modeling language vagueness in privacy policies using deep neural networks. In: 2016 AAAI Fall Symposium Series. pp. 257–263 (2016)
20. Srinath, M., Sundareswara, S.N., Giles, C.L., Wilson, S.: Privaseer: A privacy policy search engine. In: International Conference on Web Engineering. pp. 286–301. Springer (2021)
21. Wilson, S., Schaub, F., Ramanath, R., Sadeh, N., Liu, F., Smith, N.A., Liu, F.: Crowdsourcing annotations for websites' privacy policies: Can it really work? In: Proceedings of the 25th International Conference on World Wide Web. pp. 133–143 (2016)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29921-5_15