

Multi-AI Agent Oriented Privacy Policy Compliance Checking Based on Large Language Models in Mobile IoT Systems

Yangyan Zeng, Wei Liu, Anqi Wang, Chunchao Zeng, Yi Yang, Wei Liang, Member, *IEEE*, Xu Zhang, Member, *IEEE*, M. Jamal Deen, Fellow, *IEEE*, Xiaokang Zhou, Member, *IEEE*

Abstract—The proliferation of the Internet of Things (IoT) has exacerbated data privacy risks, necessitating rigorous compliance auditing of IoT companion application privacy policies. However, automated detection faces unique technical challenges, such as complex cross-paragraph dependencies and semantic conflicts between devices and applications. Traditional NLP methods lack deep reasoning capabilities, while single Large Language Models (LLMs) suffer from inherent reasoning biases and hallucination risks, failing to ensure strict legal traceability. To bridge this gap, we propose an evidence-traceable, multi-agent collaborative compliance detection framework driven by heterogeneous LLMs. First, we construct a fine-grained evaluation paradigm based on the Personal Information Protection Law (PIPL) and employ an indicator-driven Retrieval-Augmented Generation (RAG) mechanism to strictly anchor model reasoning to traceable policy fragments. Second, we design an evidence-quality-driven dynamic routing mechanism via a Task Complexity Assessment Model (TCAM), which adaptively dispatches multiple heterogeneous LLM agents based on quantitative difficulty and evidence scarcity, achieving an optimal balance between computational overhead and robustness. Furthermore, an Iterative Robust Aggregation (IRA) and consistency verification module is introduced to filter single-model biases and map identified risks to specific legal provisions. Extensive experiments demonstrate that the proposed framework achieves expert-level diagnostic precision with an average accuracy of 92.5%, a Macro-F1 of 93.3%, and a Cohen's Kappa of 0.85. Meanwhile, it accelerates the auditing efficiency by approximately 30 times compared to manual reviews, supporting high-throughput processing at 168 documents per hour. This end-to-end auditable system provides a highly scalable and interpretable regulatory technology (RegTech) solution for privacy governance in massive IoT ecosystems.

The corresponding authors are Yi Yang and Xiaokang Zhou.

Yangyan Zeng is with the Basic Research Institute of New Generation Artificial Intelligence, Xiangjiang Laboratory, Changsha 410205, China (e-mail: yangyanz0930@163.com).

Wei Liu is a member of the School of Artificial Intelligence and Advanced Computing at Hunan University of Technology and Business, Changsha 410205, China (e-mail: 2985812563@qq.com).

Anqi Wang is a member of the School of Artificial Intelligence and Advanced Computing at Hunan University of Technology and Business, Changsha 410205, China (email: 1627198867@qq.com).

Chunchao Zeng is a member of the School of Artificial Intelligence and Advanced Computing at Hunan University of Technology and Business, Changsha 410205, China (email: 2983416514@qq.com).

Yi Yang is with the School of Advanced Interdisciplinary Studies, Hunan University of Technology and Business, Changsha 410205, China (e-mail: 2698@hutb.edu.cn).

Wei Liang is with the School of Advanced Interdisciplinary Studies and School of Management Science and Engineering, Hunan University of Technology and Business, Changsha 410205, China, and also the Xiangjiang Laboratory, Changsha 410205, China (e-mail: weiliang@csu.edu.cn).

Xu Zhang is with the School of Computing Sciences, University of East Anglia, x.zhang27@uea.ac.uk.

M. Jamal Deen is with the Faculty of Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: jamal@mcmaster.ca).

Xiaokang Zhou is with the Faculty of Business Data Science, Kansai University, Osaka 565-8585, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: zhou@kansai-u.ac.jp).

Index Terms—large language model, Internet of Things, privacy policy, compliance detection, multi-agent system, retrieval-augmented generation, Personal Information Protection Law.

I. INTRODUCTION

With the continuous advancement of sensing, communication, and edge computing technologies, the Internet of Things (IoT) is evolving from simple "device connectivity" into a critical digital infrastructure empowering diverse application scenarios^[1-3]. Currently, cutting-edge technologies such as decentralized graph learning, lightweight zero-trust architectures^[4], multi-modality behavioral analysis^[5], and robust causal representation^[6] are driving the development of next-generation network security and interpretable heterogeneous IoT systems^[7]. It is projected that the number of active IoT connected devices globally will reach approximately 21.1 billion by 2025, making massive and heterogeneous data a core factor of production in the digital economy^[8]. However, while significantly enhancing social efficiency, the IoT has also drastically expanded the attack surface. Not only are IoT malware attacks experiencing exponential growth^[9], but complex threats against IoT graph neural networks are also becoming increasingly severe^[10]. More critically, privacy violations and data abuse scandals (e.g., the Cambridge Analytica incident) have emerged endlessly^[11]. Such data breaches not only severely infringe upon user privacy but also cause involved enterprises to face significant negative market returns, dealing a substantial blow to their market capitalization^[12]. Furthermore, the response strategies adopted by enterprises post-breach (e.g., apologizing or justifying) directly influence investor behavior and stock value^[13]. Mandatory breach disclosures can even force companies to curtail long-term Research and Development (R&D) investments to navigate short-term crises^[14], triggering strong threat appraisals and negative reactions from consumers^[15]. This compliance risk is pervasive across the digital ecosystem: measurement studies indicate that top Chinese websites^[16], health code applications^[17], and third-party skills within the Alexa ecosystem^[18] all exhibit a significant compliance gap between their disclosed privacy policies and actual data practices.

To counter these escalating risks, various countries and regions have successively promulgated stringent data protection regulations (e.g., GDPR and PIPL). Traditional research on automated privacy compliance detection has predominantly relied on Natural Language Processing (NLP) techniques, accumulating a wealth of systematic review literature^[19]. Typical applications include utilizing NLP to automatically check the completeness of Data Processing Agreements (DPAs)^[20] and assessing privacy policy compliance through scoping analysis^[21]. However, NLP-based approaches suffer

from distinct bottlenecks: they are highly dependent on large-scale manually annotated data, exhibit limited generalization capabilities, and fall short when processing lengthy privacy legal clauses that require complex logical deduction.

In recent years, Large Language Models (LLMs) have demonstrated exceptional natural language understanding capabilities in the legal vertical domain^[22]. LLMs have been widely applied in various high-risk scenarios, such as the compliance generation of clinical research documents^[23]. Nevertheless, for rigorous privacy compliance detection, relying solely on single LLMs presents a fatal weakness: they are highly susceptible to "hallucinations" and factual deviations. To mitigate this issue, Retrieval-Augmented Generation (RAG) technology has been introduced to provide factual grounding, achieving remarkable success in areas like construction safety information generation^[24] and the fine-tuning of automated code compliance checking for building regulations^[25]. Although RAG resolves the "information recall" problem, a standalone RAG system still exhibits critical flaws: it cannot independently handle complex compliance audits requiring multi-step logical deduction, and it is highly vulnerable to retrieval noise.

Therefore, achieving high-confidence automated compliance detection necessitates the introduction of a more advanced collaborative paradigm. Particularly in the IoT ecosystem, the automated auditing of privacy policies faces unique technical challenges: the first is "Cross-paragraph dependencies," where data collection, processing, and sharing clauses are often scattered across different sections of lengthy texts, demanding extremely strong long-range logical association capabilities from the model; the second is "Semantic conflicts," where inconsistencies or even contradictions frequently exist between hardware data collection instructions on the IoT device side and the privacy statements on the companion Application (App) side. Confronted with these highly complex scenarios, traditional single models or simple RAG frameworks often fail to strike a balance. Multi-Agent technology driven by LLMs provides a crucial bridge to close this gap. Unlike the linear reasoning of a single model, a multi-agent system can employ role assignment, multi-perspective cross-validation, and collaborative reasoning to effectively identify hidden cross-paragraph links and reconcile semantic conflicts between devices and applications. Consequently, it dismantles the complex logical chains inherent in IoT privacy compliance and fundamentally filters out the "hallucinations" and reasoning biases inherent in single models. Similar architectures have already demonstrated unparalleled advantages in tasks such as cross-validating code for the automated design of reinforced concrete structures^[26] and filtering single-model biases in multi-defendant legal judgment predictions^[27]. Combined with the factual anchors provided by RAG, this cutting-edge architecture has shown immense potential in software requirement compliance verification^[28].

To rigorously quantify the actual efficacy of the aforementioned multi-agent mechanism in mitigating LLM hallucination risks, we did not limit our approach to pure automated evaluation. Instead, we relied on rigorous manual annotation by domain experts with interdisciplinary backgrounds to construct a reliable ground truth for factual auditing. In subsequent experimental evaluations, we conduct dedicated ablation studies targeting hallucination elimination, empirically validating the framework's effectiveness by quantitatively comparing factual error rates and compliance

judgment accuracy under single-model versus multi-agent collaborative settings. Based on this, this paper proposes an automated detection framework that integrates dynamic routing and multi-agent consistency verification. The system explicitly adopts an ensemble strategy of Heterogeneous LLMs and introduces statistical optimization concepts to significantly enhance the robustness of the detection conclusions. The core contributions of this paper are summarized as follows:

(1) We propose a fine-grained, legal-indicator-guided, and evidence-traceable paradigm for IoT privacy compliance evaluation. Addressing the complex privacy scenarios of IoT companion apps, we decompose the compliance requirements of the Personal Information Protection Law (PIPL) into a computable, indicator-level evaluation system. The validity of this system has been rigorously verified through comprehensive regulatory coverage checks and consistency evaluations by domain experts, enabling the precise localization of privacy compliance vulnerabilities and repeatable engineering deployment.

(2) We construct an indicator-driven RAG evidence retrieval and factual traceability mechanism. By building a vectorized privacy knowledge base and implementing indicator-specific retrieval, we not only effectively filter contextual noise when LLMs process lengthy texts but also ensure that all compliance decisions are strictly anchored to verifiable original clauses, fundamentally curbing model hallucinations.

(3) We design an evidence-quality-driven dynamic multi-agent collaborative mechanism that adaptively balances reasoning efficiency and robustness. To address the variance in the complexity of IoT privacy clauses and evidence quality, the system explicitly defines a two-dimensional quantitative evaluation standard comprising "inherent indicator difficulty" and "evidence scarcity." Specifically, evidence scarcity is measured with fine granularity via retrieval recall rates, cross-section dispersion degrees, and semantic ambiguity. Based on this multi-dimensional difficulty metric, the system introduces a dynamic routing mechanism that adaptively dispatches multiple heterogeneous LLM agents to conduct parallel cross-evaluations for tasks of varying difficulties. This heterogeneous ensemble mechanism optimizes computational reasoning costs while effectively overcoming the attention drift and homogeneous reasoning biases of single models when processing ambiguous semantics.

(4) We build and validate an end-to-end auditable closed-loop system mapping policy texts to legal clauses. Breaking through the fragmentation limitations of traditional methods, the system not only outputs high-confidence conclusions via multi-model consistency analysis but also automatically maps identified privacy risks precisely to specific legal origins, generating structured audit reports.

The remainder of this paper is organized as follows: Section II reviews the related work. Section III introduces the proposed methodological framework and its key modules. Section IV describes the experimental setup, comparative analysis, and ablation studies. Section V concludes the paper.

II. RELATED WORK

A. Traditional NLP Methods and Privacy Compliance

In the early stages of compliance automation, deep learning and structured parsing played a central role. Chen et al.^[29] proposed VioDroid-Finder to automatically evaluate the privacy consistency of Android applications, while Tu et al.^[30]

utilized pre-trained BERT models to substantially enhance the efficiency of privacy phrase extraction and permission detection. For emerging ecosystems, Wang et al. [31] tailored taint analysis techniques to reveal the alarmingly high inconsistency rates of data practices in Mini-Apps. Concurrently, Nguyen et al. [32] developed the XLMD framework using multilingual models to resolve privacy consistency detection challenges in resource-poor languages (e.g., Vietnamese). However, as highlighted by NLP surveys in the legal domain, traditional NLP methods remain constrained by prohibitive annotation costs and face a critical bottleneck in long-range logical association capabilities when conducting cross-paragraph reasoning [33].

B. Applications of LLMs, RAG, and Multi-Agent Systems in the Legal Domain

Generative AI is reshaping complex legal and engineering review paradigms. Studies have confirmed that LLMs can effectively execute tasks ranging from interpreting BIM building regulations [34] and translating contract texts into logical predicates [35], to generating compliant smart contracts based on blockchain [36]. By integrating knowledge graphs or fine-tuning techniques, LLMs have demonstrated outstanding performance in contract question-answering [37], automated contract reviews [38], public procurement legal consulting [39], and the compilation of regulatory codes [40]. To further eliminate hallucinations, RAG technology has become a standard configuration in vertical domains. Whether through contrastive sentence generation [41], retrieval augmentation for safety risk management [42], optimization for low-resource language retrieval [43], or visual-graph RAG integrated with large multimodal models [44] and dual-stage fine-tuned reasoning [45], RAG has proven its reliability in factual traceability for complex information extraction [46] and GDPR requirement capturing [47]. The "aggregation agent" and "consistency analysis" mechanisms in multi-agent architectures are not isolated innovations; their deep technical lineage can be traced back to consensus algorithms in distributed systems (e.g., Practical Byzantine Fault Tolerance, PBFT) and Ensemble Learning voting mechanisms in classical machine learning. In a recent 2024 study, Chen et al. [48] significantly improved the robustness of software defect prediction and classification by constructing a multi-classifier diversity ensemble network. The latest cutting-edge research has begun to deeply integrate distributed consensus algorithms, such as Weighted Byzantine Fault Tolerance (WBFT), into Multiple Large Language Model Networks (Multi-LLMNs). Through the distributed collaboration and voting of heterogeneous agents, this approach effectively resists "hallucinations" and Byzantine errors generated by single models [49].

In the context of edge collaboration, Yang et al. [50] investigated task offloading in mobile edge computing using multi-agent deep reinforcement learning [51]. However, such reinforcement-learning-based routing mechanisms often entail exorbitant training overheads and lack decision transparency. In contrast, the FrugalGPT framework proposed by Chen et al. [52] for LLM inference optimization demonstrated that heuristic-rule-based LLM cascading and Early Exit mechanisms can drastically reduce computational overhead without sacrificing accuracy [53]. Building upon this foundation, modern Multi-Agent architectures go a step further to achieve system-level self-correction and knowledge verification. Dong et al. [54] demonstrated the coordination capabilities of agents for non-expert interactions in BIM collaboration, while

Sadowski et al. [55] combined modular multi-agents with symbolic reasoning, unprecedentedly enhancing the transparency and consistency of legal reasoning.

C. Applications of RAG and LLMs in Privacy Compliance

In the specific scenario of data protection and privacy compliance, a myriad of algorithmic solutions continues to emerge. Early machine learning has been successfully deployed for automated GDPR completeness checking [56] and DPA text classification [57]. More recently, advanced models (such as the Privacify system) have significantly improved cross-document summary analysis and the user readability of policies [58], while the FACTOR framework achieved compliance reasoning for healthcare services through LLM textual entailment [59]. To further overcome retrieval bottlenecks in complex compliance domains, recent top-tier research has heavily focused on advanced Re-ranking and Iterative retrieval strategies. For instance, Yue et al. [60] proposed the RankRAG framework, which unifies context re-ranking and the generation process within LLMs, substantially filtering redundant retrieval noise and boosting recall rates. Within the legal technology vertical, a recent comprehensive survey by Alkhatib et al. [61] evaluated the optimization pathways of legal RAG pipelines, confirming that the introduction of precise re-ranking mechanisms and enhanced retrieval can drastically reduce model hallucinations in legal judgments and enhance the interpretability of outputs. Additionally, RAG frameworks based on event graphs for structuring regulatory information [62], the CO2 ontology system driving AI compliance [63], and the QA-RAG dual-track mechanism [64] have all effectively optimized retrieval precision. Regarding the mitigation of sensitive data risks, the LPRAG framework, rooted in Local Differential Privacy, has provided new theoretical backing for data security in compliance detection [65].

D. IoT Privacy Compliance Detection

The IoT ecosystem faces unique privacy compliance challenges due to cross-device data flows and third-party API integrations. The fully automated assessment method by Guamán et al. [66] identified severe GDPR violation risks regarding cross-border personal data transfers in Android applications. Dynamic analyses of consumer wearable devices by Doherty et al. [67] exposed vast disparities in industry-wide data governance policies, while the IoTPrivComp measurement study further confirmed the prevalent existence of undisclosed data leaks in IoT apps [68]. To achieve more scalable detection, Oranekwu et al. [69] made preliminary attempts to combine ontology-driven reasoning with RAG for the real-time automated assessment of IoT cybersecurity guidelines. However, the most relevant state-of-the-art works mentioned above still suffer from two fundamental methodological limitations, which constitute the core differentiation of our research: On the one hand, existing detection methods predominantly rely on macroscopic regulatory mapping or qualitative analysis; they lack a fine-grained, indicator-driven system, making it exceedingly difficult to achieve the precise quantification and localization of compliance vulnerabilities. On the other hand, although some studies have incorporated basic RAG mechanisms, they are mostly confined to single-model linear reasoning and fail to employ Multi-Agent collaborative systems to tackle the complex cross-paragraph dependencies and semantic conflicts inherent in IoT texts. Therefore, regarding the highly complex and frequently contradictory privacy clauses of IoT companion apps, there remains a critical absence of a detection framework

that seamlessly integrates "precise evidence constraints of RAG" with the "deep cross-logical reasoning of multi-agents." This is precisely the core problem that this study aims to resolve.

III. MULTI-AI AGENT CONSTRUCTION

This study proposes an automated compliance detection framework for privacy policies associated with IoT device companion applications. The core architecture is shown in Fig. 1. The framework is designed to balance detection efficiency, robustness of results, and explainability, and to control the hallucination risk of large language models through a structured process.

Terminology Clarification: Before detailing the specific modules of this architecture, it is imperative to clarify the scope of the term "Agent" as utilized in this study. In mainstream artificial intelligence and reinforcement learning, an "agent" typically implies an autonomous entity with continuous perception, strategic planning, action execution, and environmental learning capabilities. In contrast, the "agents" in our framework are employed in a restricted, application-specific sense. They serve as static, one-off "LLM-based decision units" or "evaluation experts." They do not possess continuous learning capabilities or engage in dynamic negotiation and planning. Instead, they act as independent parallel processors within a heterogeneous ensemble, performing zero-shot logical reasoning under strict prompt constraints to mitigate the hallucinations and biases inherent in single models.

A. IoT Privacy Compliance Detection Framework Based on Multi-Agent

To support implementation and enable quantitative evaluation, we first collected and organized privacy policy texts from representative IoT companion applications. These texts formed the detection dataset. Guided by China's Personal Information Protection Law (PIPL), we established an evaluation indicator system with corresponding judgment criteria. This system provides a unified benchmark for the subsequent detection process.

Next, we constructed a privacy knowledge base using a retrieval-augmented generation (RAG) scheme. The policy texts were segmented into chunks, converted into vector representations, and stored in a vector database. This representation allows the system to retrieve the most relevant original passages for each compliance indicator. This step significantly reduces irrelevant text input and saves computational cost. More importantly, it ensures that all subsequent judgments are grounded in traceable original evidence.

To optimize computational resources, the system introduces a dynamic routing and scheduling mechanism driven by a Task Complexity Assessment Model (TCAM). This module quantitatively assesses the retrieved evidence based on its semantic clarity, information completeness, and inherent legal difficulty. Based on the resulting complexity score, the system adaptively allocates its inference budget by dispatching either a single model for straightforward text or an ensemble of three to five heterogeneous models for cases involving ambiguous or fragmented clauses. This adaptive multi-agent collaboration ensures a precise balance between automated audit efficiency and reasoning accuracy.

During the compliance judgment stage, the system uses either binary decisions (for example, compliant or non-compliant) or fine-grained scores, depending on the nature of

each indicator. Each result is accompanied by detailed reasoning to support further analysis. The results are then aggregated by voting for binary decisions and by weighted aggregation for scores. A consistency check is performed on the models' reasoning to identify points of consensus and disagreement, which improves the reliability of the final conclusions.

Finally, for indicators that remain risky after aggregation, the system automatically retrieves the relevant legal provisions. This step aligns the technical judgment with the legal basis. A structured compliance report is then generated. The report presents the complete audit chain, including the detection findings, source text evidence, legal citations, and rectification suggestions. In this way, it forms an end-to-end evaluation loop.

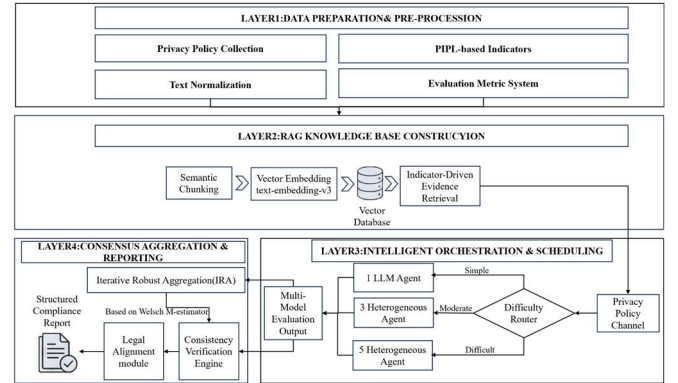


Fig. 1. Privacy agreement compliance detection system architecture diagram

B. Dataset and Evaluation Indicator System Construction

Dataset construction: We constructed a corpus of privacy policies for IoT companion mobile applications in a specific domain. The corpus was manually collected to reflect real-world deployments and compliance practices in heterogeneous IoT scenarios. The dataset consists of 100 privacy policies, classified into three representative categories: smart monitoring applications (27 policies), dashcam applications (37 policies), and smart home applications (36 policies). These categories were chosen because they typically involve continuous sensing, cross-device data flows, and frequent third-party SDK integrations, which increase the complexity of compliance and auditing.

We collected the policies by manually reviewing policy pages from official distribution channels and vendor websites to ensure that each document corresponds to an actually deployed application. For each application, we removed duplicates, for example the same policy used in different applications without version differences, and non-policy documents, for example user agreements without personal data clauses. When multiple policy pages existed for the same application, for example a web version and an in-app version, we retained the version with more complete and more consistent disclosures.

Construction of the evaluation indicator system: To ensure that the criteria for compliance testing are legally traceable and reproducible in engineering practice, we constructed an evaluation indicator system based on the core obligations and audit concerns of the Personal Information Protection Law (PIPL). To ensure the scientific rigor, effectiveness, and legal soundness of the assessment indicator system, this study adopted a rigorous verification mechanism. First, we conducted a comprehensive coverage check, precisely

mapping the 40 extracted secondary indicators to specific provisions of the Personal Information Protection Law (PIPL). For example, indicators B1 "Privacy Policy Name" and B5 "Data Processor Identity" directly correspond to Article 17 of the PIPL (requirements for transparency in personal information processing rules and the authenticity of processor identities); while B13 "Minimum Necessity Principle" strictly corresponds to Article 6 of the PIPL. Second, to verify the rationality of these indicators in practical IoT privacy audits, we introduced the Delphi method, inviting five senior experts with interdisciplinary backgrounds in law and computer science to conduct independent evaluations. The expert group conducted multiple rounds of blind review and scoring of the necessity, clarity, and operability of each secondary indicator.

Subsequently, we used Fleiss's kappa coefficient to conduct a statistical test on the consistency of the expert evaluation. The results showed that the Kappa values of all indicators were highly consistent ($kappa > 0.80$). The system consists of seven primary indicators and forty secondary indicators, as shown in Table I.

In this system, secondary indicators are used as the smallest unit of judgment. For each secondary indicator, the system specifies the necessary disclosure elements and verifiable evidence anchors and produces structured outputs in a unified format during both the annotation and model inference stages. These outputs support subsequent quantitative assessment, result aggregation, and audit report generation.

Table I Privacy Policy Completeness Assessment Indicators and Selection Rationale

Primary indicators	Symbol	Secondary indicators	Symbol	Weight	Indicator basis
Basic Information	A_1	Privacy Policy Name	B_1	1.79%	Article 17、22、52 of the PIPL
		Privacy Policy Version Number	B_2	1.66%	
		Privacy Policy Effective Date	B_3	2.04%	
		Privacy Policy Update Date	B_4	1.63%	
		Privacy Policy Responsible Person	B_5	2.19%	
		Privacy Policy & Contact Information	B_6	2.43%	
	Data Collection Rules	Purpose and Scope of Personal Information Collection	B_7	2.72%	Articles 13, 14, 30, and 31 of the PIPL
		Transparency in the Collection of Personal Information	B_8	2.08%	
		Cookies or similar technologies	B_9	1.90%	
		Notification and Consent Mechanism offers options	B_{10}	2.81%	
		Protection of Minors' Personal Information	B_{11}	2.71%	
			B_{12}	2.77%	
Data Usage Rules	A_3	Minimum Necessity Principle	B_{13}	2.62%	Articles 6, 13, 28, and 29 of the PIPL
		Purpose Restriction	B_{14}	2.59%	
		Special Purpose Usage Rules	B_{15}	2.66%	
Data Storage and Security Measures	A_4	Personal Information Storage Location	B_{16}	2.60%	Articles 6, 47, 51, and 53 of the PIPL
		Expected Retention Time for Personal Information	B_{17}	2.63%	
		Safety Qualification Description	B_{18}	1.72%	
		Security Measures	B_{19}	2.71%	
		Emergency handling of personal information leaks	B_{20}	2.64%	
			B_{21}	2.42%	
		Personal Information Security Self-Inspection and Assessment	B_{22}	2.70%	
		Personal Information Security Incident Notification			
		Data Access Rights	B_{23}	2.72%	
		Right to Correction, Right to Deletion	B_{24}	2.71%	
Personal Information and Subject Rights	A_5	Withdrawal of Consent	B_{25}	3.73%	Articles 44-48、15, and 50 of the PIPL
		Data migration and download service	B_{26}	2.52%	
		Privacy Preferences Settings	B_{27}	2.17%	
		Right to Cancel Account	B_{28}	2.70%	
		Supervision and Complaint Methods	B_{29}	2.58%	
		Clearly defines the scope, purpose, method, and timeframe.	B_{30}	2.80%	
		User's explicit authorization	B_{31}	2.78%	
		Recipient Identity	B_{32}	2.72%	
Rules for Entrusted Processing, Sharing, Transfer, and Public Disclosure	A_6	Rights and Obligations of the Recipient	B_{33}	2.63%	Articles 23, 25, and 38 of the PIPL
		Public Disclosure of Risks	B_{34}	2.70%	
		Cross-border Data Transfer	B_{35}	2.63%	

Children's Privacy	A_7	Collection of Children's Personal Information	B_{36}	2.76%	Article 31 of the PIPL
		Child Identity Smart Trigger	B_{37}	2.76%	
		Anonymization	B_{38}	2.73%	
		Child Privacy Protection	B_{39}	2.63%	
		Security Risk Assessment	B_{40}	2.54%	

To scientifically quantify the relative importance of the 40 secondary indicators, we formalize the weight assignment as a multi-dimensional feature mapping problem. We propose a tripartite weighting model that integrates "regulatory intensity, empirical distribution, and expert scenario risk." Specifically, we define the indicator feature matrix as $X \in \mathbb{R}^{S \times 3}$, where $S = 40$ is the total number of secondary indicators. Each row vector $x_{ij} = [L_{ij}, F_{ij}, V_{ij}]$ represents the compliance characteristics of the j -th secondary indicator under the i -th primary indicator across three dimensions:

1) Legal and Normative Intensity (L_{ij}): This dimension serves as the regulatory baseline. $L_{ij} \in (0, 1.0]$ characterizes the mandatory status of the obligation within the current legal hierarchy. Indicators directly corresponding to the core provisions of the Personal Information Protection Law (PIPL) are assigned the highest regulatory intensity, ensuring the evaluation remains anchored to national statutory requirements.

2) Empirical Disclosure Frequency (F_{ij}): To enhance the statistical robustness of the empirical weights and mitigate bias from niche phrasing, we expanded our analysis to a large-scale, cross-industry dataset comprising 583 diverse privacy policies. By constructing an indicator-specific dictionary, the system executes fine-grained text scanning to calculate the empirical disclosure ratio. The frequency is formally defined as:

$$F_{ij} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathbb{I}(\text{hit}_{ij,d}) \quad (1)$$

where $|\mathcal{D}| = 583$ is the total number of documents in the corpus, and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the core semantics of indicator B_{ij} are detected in document d , and 0 otherwise. This large-scale cross-industry baseline allows the system to filter out idiosyncratic noise while precisely identifying "high-risk, low-disclosure" vulnerabilities.

3) Expert Risk Assessment (V_{ij}): Considering the unique vulnerabilities of IoT ecosystems (e.g., cross-device data fusion and passive sensing), a panel of interdisciplinary experts evaluates the actual risk exposure $V_{ij} \in (0, 1.0]$ for each indicator. This compensates for the inherent latency of legal provisions in perceiving evolving privacy engineering practices.

4) Weight Fusion Operator:

To map the feature matrix X to the final weight distribution, we propose a normalized linear combination strategy. To balance the dominance of high-frequency boilerplate language, a non-linear smoothing function $\Phi(x) = \ln(1 + x)$ is applied to the empirical frequency component. The final composite weight W_{ij} is formulated as:

$$W_{ij} = \frac{\alpha L_{ij} + \beta \Phi(F_{ij}) + \gamma V_{ij}}{\sum_{k=1}^M \sum_{l=1}^{N_k} (\alpha L_{kl} + \beta \Phi(F_{kl}) + \gamma V_{kl})} \quad (2)$$

where $M = 7$ denotes the number of primary indicators, and

N_k represents the number of secondary indicators within the k -th category. The fusion coefficients are set to $\alpha = 0.5, \beta = 0.2, \gamma = 0.3$, strictly enforcing the "Regulatory Priority, Risk Adjustment, and Empirical Correction" design principle.

5) Modeling Results and Computational Complexity:

Through this mapping operator, the discrete textual compliance logic is successfully transformed into a computable quantitative benchmark. Table II illustrates the modeling parameters for representative indicators. (The complete 40×4 weight matrix is detailed in Appendix A). The proposed weighting model is highly lightweight. Computing the composite weight W_{ij} requires only $\mathcal{O}(1)$ constant time per secondary indicator. Consequently, generating the complete weight distribution for the entire system takes strictly linear time $\mathcal{O}(S)$ (where $S = 40$), with a minimal space complexity of $\mathcal{O}(S)$ for storing the weight matrix. This negligible overhead ensures that the weight allocation module scales effortlessly to process massive heterogeneous IoT applications, perfectly satisfying the efficiency and dynamic routing requirements of our multi-agent architecture without becoming a computational bottleneck.

Table II: Weighting Model Parameters for Representative Secondary Indicators

ID	Indicator Name	L_{ij}	F_{ij}	V_{ij}	W_{ij}
B1	Privacy Policy Name	0.70	89.71%	0.40	1.79%
B7	Purpose and Scope of Collection	1.00	99.31%	0.80	2.72%
B13	Minimum Necessity Principle	0.70	84.73%	0.80	2.62%
B26	Data Portability Services	0.40	21.09%	0.40	2.52%

C. Indicator-Driven Evidence Retrieval

The retrieval-augmented generation (RAG) knowledge base is designed to transform lengthy and unstructured privacy policies into a searchable evidence space for downstream compliance reasoning. By constraining model inference to retrieved policy fragments, the proposed module reduces irrelevant context, lowers token consumption, and improves the traceability and factual grounding of subsequent judgments. The construction and retrieval procedure is organized into the following steps:

Step 1: Text Normalization and Semantic Chunking

Given a raw privacy policy document D^{raw} , the system first performs text normalization to obtain:

$$D = \Phi(D^{raw}) \quad (3)$$

where $\Phi(\cdot)$ replaces consecutive spaces, tab characters, and line breaks with a single blank character. Let the normalized document be denoted as:

$$D = x_1, x_2, \dots, x_{|D|} \quad (4)$$

where x_t is the t -th character and $|D|$ is the document length.

To preserve local semantic continuity while controlling fragment granularity, a fixed-length sliding-window chunking strategy is adopted. Let $L_{\max}$ denote the chunk length and L_{ov} denote the overlap length. The stride is defined as:

$$\Delta = L_{\max} - L_{ov} \quad (5)$$

The j -th chunk is then constructed as:

$$c_j = D[(j-1)\Delta + 1 : \min((j-1)\Delta + L_{\max}, |D|)] \quad (6)$$

where $j = 1, 2, \dots, N$, and:

$$N = \left\lceil \frac{|D| - L_{ov}}{L_{\max} - L_{ov}} \right\rceil \quad (7)$$

Thus, the document is transformed into a chunk set:

$$\mathcal{C}(D) = c_1, c_2, \dots, c_N \quad (8)$$

In this work, $L_{\max} = 256$ and $L_{ov} = 30$. Such an overlap design helps retain cross-boundary semantics for long legal clauses, especially those involving third-party sharing, retention rules, or multi-stage data processing.

Step 2: Vector Embedding and Dual-Index Construction

Each chunk $c_j \in \mathcal{C}(D)$ is mapped into a dense semantic vector through the embedding model:

$$v_j = f_{emb}(c_j) \in \mathbb{R}^d \quad (9)$$

where $f_{emb}(\cdot)$ denotes the selected embedding function and d is the embedding dimension. The chunked document is therefore represented as:

$$\mathcal{V}(D) = \{(c_j, v_j)\}_{j=1}^N \quad (10)$$

The vectorized chunks are then stored in Elasticsearch to form the retrieval knowledge base. Formally, the indexing operation can be written as:

$$\Psi: \mathcal{V}(D) \rightarrow \mathcal{K} \quad (11)$$

where $\mathcal{K}$ denotes the searchable knowledge base. In implementation, $\mathcal{K}$ combines HNSW-based approximate nearest neighbor indexing with the native inverted index, enabling both semantic retrieval and keyword-aware matching. This hybrid structure improves robustness when privacy-policy expressions vary across applications and categories.

Step 3: Indicator-Driven High-Quality Hybrid Retrieval

During the retrieval phase, the system deploys an indicator-driven query mechanism to achieve precise evidence recall. Relying on the constructed dual-index, the system executes a high-quality hybrid retrieval strategy to enhance the robustness of complex legal text retrieval. For a specific primary compliance indicator A_i , the system automatically generates a structured semantic query set $Q_i = \{q_{i1}, \dots, q_{it}\}$, covering various paraphrases and expressions that the indicator might assume within privacy policies. Each query q_{is} is projected into the vector space to obtain the query vector $e_{is} = f_{emb}(q_{is})$.

Under the hybrid retrieval framework, the core semantic similarity maximization recall process is formalized as:

$$\mathcal{E}_i = \bigcup_{s=1}^t \left\{ c_j \mid \frac{e_{is}^{(s)} \cdot v_j}{\|e_{is}^{(s)}\| \|v_j\|} \geq \tau_s, j \in \text{Top-}k \right\} \quad (12)$$

where τ_s is a dynamic similarity threshold, and Top- k restricts the output to the k candidate fragments with the

highest similarity scores. This cosine similarity metric effectively captures the semantic relevance between the query and document fragments, overcoming the limitations of traditional keyword matching.

To balance global precision and local evidence completeness, the dynamic threshold τ_s is determined by a hybrid strategy considering both indicator sensitivity and the statistical score distribution of the candidate set:

$$\tau_s = \tau_i - \lambda \cdot \sigma_s \quad (13)$$

where τ_i is the baseline threshold customized for primary indicator A_i , and σ_s represents the standard deviation of similarity scores in the current candidate set, with a scaling factor empirically set to $\lambda = 0.5$. For rigid indicators such as A_1 (Basic Information), a strict threshold $\tau_i = 0.65$ is established to filter irrelevant contextual noise. Conversely, for indicators modeling complex data flows, such as A_6 (Sharing & Transfer) and A_7 (Children's Privacy), a lower threshold $\tau_i = 0.45$ is adopted to maximize the recall of scattered evidence. This adaptive strategy ensures that the system maintains a high recall rate for sensitive privacy clauses while relying on the subsequent re-ranking model to prune noise.

Step 4: Lightweight Re-ranking and Refinement

To further refine the retrieved candidate set $\mathcal{E}_i$, a lightweight reranking model is introduced. For each query-fragment pair (q_{is}, c_j) , the reranking score is computed as:

$$R(q_{is}, c_j) = f_{rerank}(q_{is}, c_j), \forall c_j \in \mathcal{E}_i \quad (14)$$

where $f_{rerank}(\cdot, \cdot)$ denotes the reranking function. If a fragment is matched by multiple query variants, its final score is aggregated as:

$$\hat{R}(c_j) = \max_{1 \leq s \leq t} R(q_{is}, c_j) \quad (15)$$

The final evidence sequence is obtained by sorting all candidates in descending order of $\hat{R}(c_j)$:

$$\hat{\mathcal{E}}_i = \text{Rank}(\mathcal{E}_i; \hat{R}(c_j)) \quad (16)$$

This reranking stage improves evidence quality by emphasizing clause-level semantic alignment and information completeness, thereby providing a more reliable and traceable evidence package for downstream compliance evaluation.

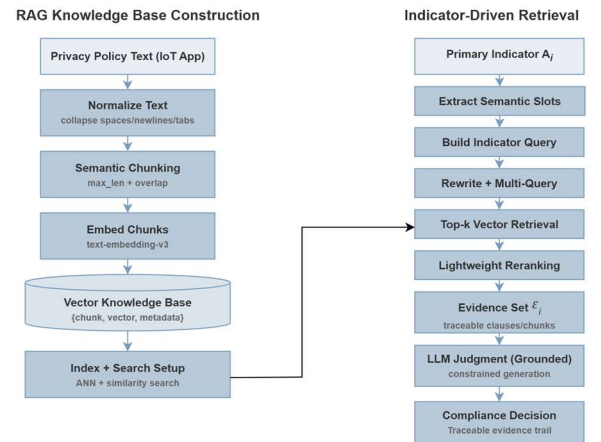


Fig. 2. RAG knowledge base construction and original text retrieval

D. Dynamic Routing and Scheduling

Privacy policy text varies substantially in writing style and information density. The same compliance indicator may

appear in different applications as a single clear statement, as descriptions scattered across multiple paragraphs, through vague terms (including but not limited to...), or within interrelated clauses. Applying a fixed retrieval and inference configuration to all indicators not only leads to unnecessary computational cost but also increases the risk of model speculation when the evidence is insufficient. Therefore, we propose an evidence-quality-driven dynamic routing and scheduling mechanism. This mechanism automatically selects an appropriate inference budget and combination of models for each indicator in accordance with the current policy, as illustrated in Fig. 3.

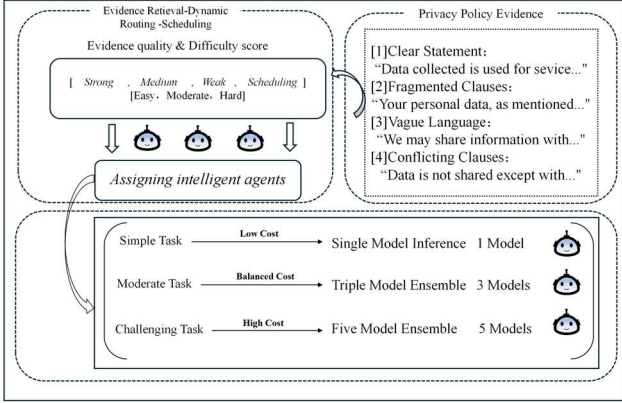


Fig. 3. Schematic diagram of dynamic routing and scheduling mechanism

For the candidate evidence set ε_i of the i -th primary indicator A_i (obtained from Top-k retrieval followed by reranking), the router no longer determines the difficulty level through qualitative rules such as “Simple,” “Moderate,” or “difficult.” Instead, to guarantee objectivity, reproducibility, and quantitative interpretability, we introduce a Task Complexity Assessment Model (TCAM) to formalize the routing process. It must be clearly stated that TCAM is not a pre-trained machine learning classifier, but rather a deterministic, heuristically based mathematical model. It directly computes routing signals through a linear combination of inherent legal difficulty priors and real-time retrieval metrics such as recall, dispersion, and semantic ambiguity, thus enabling zero-shot, training-free deployment. This completely eliminates the unpredictable random errors inherent in learning-based routing models, ensuring the mathematical accuracy of difficulty judgments and enabling reproducibility based on the state of evidence.

For each primary indicator A_i , the comprehensive complexity score is computed as:

$$\theta_i = w_h H_i + w_s S_i \quad (17)$$

Here, w_h and w_s represent the weights of the inherent difficulty term and the evidence scarcity term, respectively, satisfying $w_h + w_s = 1$, $H_i \in [0,1]$ denotes the intrinsic difficulty prior of A_i derived from legal-logical modeling. Disclosure-oriented indicators such as A_1 are assigned relatively low initial difficulty values, whereas indicators such as A_6 and A_7 are assigned higher priors because they usually involve cross-entity data flows, highly sensitive information, and stronger contextual dependence.

The evidence scarcity term is computed online from the RAG-retrieved fragments as:

$$S_i = 1 - (\beta_1 R_{(ev)} + \beta_2 (1 - D_{(ev)}) + \beta_3 (1 - A_{(ev)})) \quad (18)$$

In this expression, $R_{(ev)}$ denotes the evidence recall rate,

namely the hit ratio of key compliance entities required by the corresponding secondary indicators; $D_{(ev)}$ denotes the cross-section dispersion of the retrieved evidence in the original policy; and $A_{(ev)}$ denotes the semantic ambiguity estimated from embedding-space variance. The coefficients β_1 , β_2 , and β_3 satisfy $\beta_1 + \beta_2 + \beta_3 = 1$. A larger S_i indicates that the current evidence is sparser, more fragmented, or more semantically ambiguous, and therefore implies a higher routing difficulty.

Accordingly, the scheduling decision for A_i is strictly formalized as follows: one evaluation agent is activated when $\theta_i < 0.4$; three evaluation agents are activated when $0.4 \leq \theta_i < 0.7$; and five evaluation agents are activated when $\theta_i \geq 0.7$.

Therefore, the inference budget is determined by the fused score θ_i rather than by subjective verbal boundaries. This quantitative mechanism integrates expert knowledge of legal complexity with real-time evidence quality analysis, thereby eliminating the error propagation risk caused by subjective routing failure. Specifically, this mathematically hard threshold design avoids the risk of serious misjudgments that could otherwise render subsequent evaluation processes invalid. As long as the retrieved evidence is objectively sparse, scattered throughout paragraphs, or semantically ambiguous, the sparsity term (S_i) will inevitably push the overall complexity score (θ_i) above the high-difficulty threshold. This ensures that complex or uncertain statements strictly trigger multi-agent integration, thus structurally preventing any single agent from independently processing highly uncertain contexts. The router still outputs the evidence status for each secondary indicator together with short traceable quotations, but the final difficulty label is now obtained by thresholding θ_i under TCAM, which makes the downstream resource allocation scientifically grounded and stable.

At the implementation level, to address the limited size of training corpora, reduce hallucinations in large models, and ensure the authenticity and accuracy of responses, we treat prompt engineering as part of the system methodology rather than merely as an implementation detail. Through structured design and constrained expression of prompt templates, we explicitly define “what the model can answer, what it must rely on, and how it should output” without using additional training data. This design improves the authenticity, stability, and verifiability of the inference results.

Specifically, the system closely integrates prompt engineering with dynamic routing and scheduling to construct an auditable control loop. The core mechanism of this loop is as follows. For each indicator to be inspected, the system calls a routing and scheduling node. This node operates strictly on the retrieved evidence set, that is, it fully follows the prompt word templates in Table I, and outputs a structured JSON object that contains the evidence audit results and the difficulty determination.

The generation of this JSON object strictly follows the role and task specifications defined in Table III. First, for each secondary indicator, the router evaluates the evidence status as *strong*, *medium*, *weak*, or *missing* based solely on the retrieved evidence set, and simultaneously outputs a short quotation from the original policy text as a traceable citation anchor. These short quotations play a critical role in the framework: they not only expose missing elements, vague wording, and potential contradictions directly at the source level, but also preserve the minimal yet locatable textual basis required for downstream verification, cross-model comparison, and post-hoc auditing. In this way, uncertainty is not hidden behind an abstract label, but

is instead grounded in auditable fragments of the original document. Second, rather than assigning the overall difficulty level through qualitative prompt rules such as “most” or “many” secondary indicators being sufficiently supported, the router aggregates the secondary-indicator evidence signals under the proposed Task Complexity Assessment Model (TCAM) and computes the quantitative complexity score θ_i . The final difficulty level is then determined by thresholding θ_i , yielding the corresponding routing state of *easy*, *moderate*, or *difficult* in a formally reproducible manner. This structured JSON output, rather than a direct compliance conclusion, is used as the downstream scheduling signal. Specifically, when the

routing result is *easy* ($\theta_i < 0.4$), only one evaluation model is activated for lightweight assessment; when it is *moderate* ($0.4 \leq \theta_i < 0.7$), three evaluation models are activated in parallel to balance robustness and computational overhead; and when it is *difficult* ($\theta_i \geq 0.7$), five evaluation models are jointly scheduled to better handle sparse evidence, ambiguous wording, cross-paragraph dependencies, and potential clause conflicts. By tightly coupling prompt engineering with quantitative routing, the system transforms evidence auditing, citation grounding, and inference-budget allocation into a unified and auditable control process.

Table III Design of Prompt

Role	You are the "routing agent and evidence auditor" for privacy compliance audits. Based solely on the retrieved search fragments, assess the evidence status of each secondary indicator of the current metric, extract short traceable quotations from the original text whenever available, and generate a structured routing signal for downstream scheduling under the quantitative task complexity assessment mechanism.
Task	#Judgment Rules: -If no direct evidence (keywords/entities/explicit sentences) is found for a secondary indicator in the search segment, then its `evidence_status` is "missing". -If only general statements are provided, lacking key elements such as subject/purpose/scope/path/timeframe, then `evidence_status` is "weak". -If there are explicit actionable elements and short quotations from the original text, then `evidence_status` is "strong" or "medium" (for completeness). #Difficulty is rated in three levels: - First, aggregate the evidence condition of all secondary indicators under the current primary indicator and compute the task complexity score θ under TCAM by combining evidence support, intrinsic difficulty prior, and evidence sparsity signals. - "Easy": $\theta_i < 0.4$. - "Moderate": $0.4 \leq \theta_i < 0.7$. - "Difficult": $\theta_i \geq 0.7$. -The difficulty label must be assigned according to the quantitative routing thresholds above, rather than qualitative expressions such as “most,” “several,” or “many.”
JSON output	# Output results must strictly follow the following format: <pre>{ "difficulty": "easy/moderate/difficult", "evidence_status": [{"index": 1, "item": "secondary indicator 1", "status": "strong/medium/weak/missing", "quote": "original short quote (empty if not found)"}, {"index": 2, "item": "secondary indicator 2", "status": "strong/medium/weak/missing", "quote": "..."}] }</pre>
Output Example	Output: <pre>"text": "{ \"difficulty\": \"moderate\", \"evidence_status\": [{\"index\": 1, \"item\": \"Data access right\", \"status\": \"missing\", \"quote\": \"\"}, {\"index\": 2, \"item\": \"Right to correct, right to erasure\", \"status\": \"medium\", \"quote\": \"If we decide to respond to your deletion request, we will also notify the entities that have obtained your personal information from us, requiring them to delete it promptly, unless otherwise stipulated by laws and regulations, or these entities have obtained your independent authorization.\"}, {\"index\": 3, ...} ,...] }</pre>

E. Multi-Model Evaluation

After completing original text retrieval and dynamic routing for the indicators, the system enters a multi-model parallel evaluation stage under prompt engineering constraints. In this stage, the evidence status and difficulty level output by the router serve as scheduling signals, and the scheduler adaptively activates 1, 3, or 5 evaluation models. When the evidence elements are complete and the expression is clear, a small number of models can produce stable conclusions. When missing elements, ambiguous wording, cross-paragraph dependencies, or potential clause conflicts are detected, more heterogeneous large models from different model families and inference preferences are introduced for parallel judgment. Through "cross-perspectives," the systemic biases common to single models in privacy compliance tasks, such as benevolent

interpretation of ambiguous wording, automatic completion of missing clauses, and attention drift in long-text scenarios, are reduced. In this way, uncertainty is manifested as auditable divergence signals rather than masked by the fluent expression of a single model. The entire process is illustrated in Fig. 4.

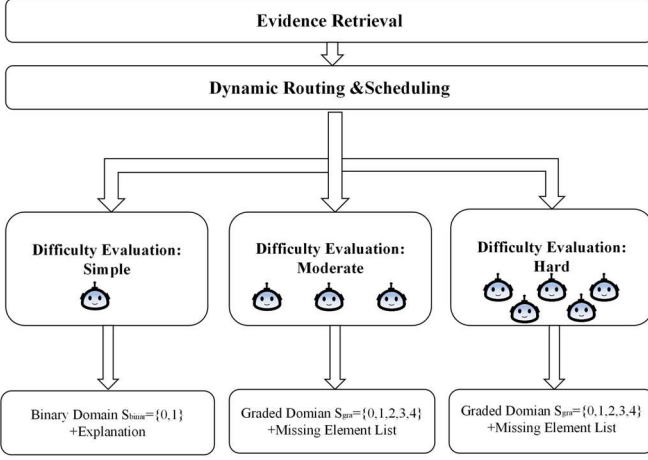


Fig. 4. Multi-model evaluation diagram

For scoring domain specification, a binary domain $\{0,1\}$ is used for disclosure-type rigid second-level indicators: $s_{r,j} = 1$ indicates the presence of a clear, locatable, and disclosure statement that satisfies the necessary elements within the evidence fragment $\mathcal{E}_i$; $s_{r,j} = 0$ indicates that the item is not mentioned or the evidence is insufficient to constitute a valid disclosure. To avoid the ambiguity of "not mentioned being misjudged as compliant" or "0 being misinterpreted as a negative compliance conclusion," the prompt template mandates that when outputting $s_{r,j} = 0$, the model must simultaneously provide a verifiable reason code $\text{reason}_{r,j}$ (e.g., not mentioned / missing elements / unclear wording) and evidence coverage explanation, ensuring that "0" consistently corresponds to a conservative stance of "insufficient or missing at the evidence level."

For non-rigid second-level indicators requiring assessment of disclosure sufficiency and executability, an ordinal scale $\{0,1,2,3,4\}$ is adopted, with fixed semantic boundaries for each level defined in the template: high scores correspond to statements with complete elements and specific, clear wording; low scores correspond to missing key elements, principle-based descriptions, or reliance on contextual inference, with 0 serving as the conservative anchor for "not mentioned or severely deficient." The model is also mandated to output a missing element list $\text{missing}_{r,j}$, ensuring that each score reduction can be explained at the evidence level, thereby reliably distinguishing between "briefly mentioned" and "executable disclosure."

More importantly, in our framework, prompt engineering is not limited to output normalization, but functions as a boundary-calibration mechanism for heterogeneous models. Specifically, all evaluation agents are anchored to the same evidence package, the same indicator definition, and the same score semantics, so that each score is interpreted against identical legal and textual criteria rather than against model-specific latent preferences. For binary indicators, the prompt explicitly defines the positive decision as requiring clear, locatable, and element-complete disclosure, while all other cases, including omission, vague wording, or incomplete elements, are conservatively mapped to 0 together with standardized reason codes. For ordinal indicators, the prompt further hard-codes monotonic score anchors by binding each level to fixed evidence conditions, completeness requirements, and executability thresholds, and simultaneously requires the model to output a missing-element list explaining why a higher score is not warranted. As a result, the prompt constrains

different model families to project their internal judgments onto a shared external rubric, thereby reducing inter-model scale drift and preventing situations in which a "3" produced by one model is semantically equivalent to a "2" produced by another. In this sense, prompt engineering serves as a lightweight calibration layer that aligns scoring boundaries before aggregation, ensuring that subsequent voting, robust fusion, and consistency verification operate on outputs that are not only structurally compatible but also semantically commensurate.

F. Aggregation and Consistency Test

Aggregation Algorithm Design: After completing the multi-model evaluation, the system obtains parallel outputs from heterogeneous models for each IoT-related application's privacy policy and each query for a target indicator (corresponding to the recall evidence set). To transform these "parallel judgments" into auditable, verifiable, and robust final conclusions, we introduce a closed-loop Aggregation and Consistency Verification module into the workflow. This module first robustly aggregates the scores and then verifies the consistency of the reasons to generate a conflict list. Suppose that the k -th primary indicator A_k triggers R independent evaluations. In practice, each evaluation outputs a structured object that contains a secondary-indicator score vector and the associated reason text. The aggregation module first performs format cleaning and input alignment on these outputs. For items not covered in a given output, the system pads the corresponding entries with 0 (equivalent to "failed to provide compliance evidence/conclusion for this item"), following a conservative strategy that avoids optimistic bias. After alignment, the score of the j -th secondary indicator $A_{k,j}$ in the r -th evaluation is denoted by $y_{k,j}^{(r)}$.

For secondary indicators whose scoring domain is the binary set $\Omega_{\text{bin}} = \{0,1\}$ (typically "basic situation" verification items), we adopt a strict majority-voting scheme. $y_{k,j}^{(r)} = 1$ indicates "compliant" and $y_{k,j}^{(r)} = 0$ indicates "non-compliant or insufficient evidence". The empirical compliance ratio is defined as:

$$p_{k,j} = \frac{1}{R} \sum_{r=1}^R y_{k,j}^{(r)} \quad (19)$$

The aggregated binary decision $\hat{y}_{k,j}$ is then obtained by thresholding the compliance ratio at one half:

$$\hat{y}_{k,j} = \begin{cases} 1, & p_{k,j} > \frac{1}{2}, \\ 0, & p_{k,j} \leq \frac{1}{2}, \end{cases} \quad (20)$$

that is, the system outputs "compliant" only when more than half of the models judge the item to be compliant, and maps ties ($p_{k,j} = 0.5$) to the non-compliant side as a conservative strategy.

To quantify the confidence of the binary decision, we characterize each sub-indicator by its "distance from 0.5" and linearly map this distance to $[0,100]$:

$$\text{CS}_{k,j}^{\text{bin}} = 200 \left| p_{k,j} - \frac{1}{2} \right| \in [0,100] \quad (21)$$

A larger $\text{CS}_{k,j}^{\text{bin}}$ implies more concentrated voting and hence higher confidence, whereas values close to zero indicate strong disagreement among models and a low-confidence judgment. It is important to acknowledge that mapping ties ($p_{k,j} = 0.5$) to the

non-compliant side introduces a deliberate conservative bias, which may systematically lower the overall compliance scores. However, in the context of regulatory technology (RegTech) and privacy compliance, a false positive (falsely judging a non-compliant or ambiguous clause as compliant) carries significant legal and security risks. Conversely, a false negative (flagging a compliant but vaguely worded clause as a risk) merely prompts further review. Compared to alternative strategies—such as treating ties strictly as an "uncertain" state that halts the automated pipeline to require mandatory human intervention—our design maintains the end-to-end fluency of the automated framework. Moreover, our system inherently accommodates the "human-in-the-loop" alternative: when a tie occurs, the calculated Confidence Score drops to zero ($CS_{k,j}^{bin} = 0$). In practical deployments, organizations can set a confidence threshold where any evaluation yielding $CS_{k,j}^{bin} = 0$ is systematically routed to a human expert for secondary review, effectively balancing the efficiency of automated auditing with the safety of human oversight.

The pseudocode for the aggregation process is as follows:

Table IV The Pseudocode for The Binary Aggregation Process

Algorithm 1: Binary Aggregation	
// aligned, missing filled with 0	
Input : scores[r][j] for r=1..R, j=1..J	
Output: s_hat[j], conf[j]	
for j = 1..J do	
sum1 = 0	
for r = 1..R do	
sum1 = sum1 + scores[r][j] // each $\in \{0,1\}$	
end for	
p = sum1 / R // compliance ratio	
if p > 0.5 then	
s_hat[j] = 1	
else	
s_hat[j] = 0 // includes tie p=0.5	
end if	
// larger distance => higher confidence	
conf[j] = LinearMap(p - 0.5 $\rightarrow$ [0,100])	
end for	
return s_hat, conf	

For sub-indicators characterized by ordered discrete grading (e.g., a 0–4 scale), simple arithmetic averaging is susceptible to perturbations caused by stochastic model hallucinations or extreme outlier evaluations. To address this, we propose an Iterative Robust Aggregation (IRA) mechanism rooted in statistical M-estimation theory. Instead of a static weighted average, we model the consensus score $y_{k,j}^*$ as the optimal solution that minimizes a global robust loss function, specifically employing the Welsch M-estimator to suppress the influence of divergent scores.

Let $y_{k,j}^{(r)}$ denote the score of the r -th evaluation round for the j -th secondary indicators, where $r \in \{1, \dots, R\}$. The aggregation is formulated as an optimization problem:

$$y_{k,j}^* = \arg \min_{\mu} \sum_{r=1}^R \rho_{\sigma}(y_{k,j}^{(r)} - \mu) \quad (22)$$

where $\rho_{\sigma}(\cdot)$ represents the robust loss function. We solve this optimization using the Iteratively Reweighted Least Squares (IRLS) algorithm. The consensus score is updated iteratively via the following fixed-point iteration rule:

$$y^{(t+1)} = \frac{\sum_{r=1}^R \omega_r^{(t)} \cdot y_{k,j}^{(r)}}{\sum_{r=1}^R \omega_r^{(t)}} \quad (23)$$

In each iteration t , the adaptive weight $\omega_r^{(t)}$ is recalculated based on the deviation of the r -th observation from the current estimated consensus $y^{(t)}$:

$$\omega_r^{(t)} = \exp\left(-\frac{(y_{k,j}^{(r)} - y^{(t)})^2}{2\sigma^2}\right) \quad (24)$$

Here, σ is the kernel bandwidth parameter that controls the sensitivity to outliers. The iteration is initialized with the median of the scores ($y^{(0)} = \text{median}(y_{k,j})$) to ensure initial robustness and typically converges within 3–5 iterations.

Finally, the converged continuous value is rounded to the nearest integer to obtain the discrete grading. To facilitate visualization and cross-indicator comparison, the system linearly maps the 0–4 rating to a 0–100 scale. A Confidence Score (CS) is derived from the convergence consistency, defined as the normalized sum of the final weights:

$$CS_{k,j} = \frac{100}{R} \sum_{r=1}^R \omega_r^{(final)} \quad (25)$$

A higher CS indicates a high degree of consensus among evaluation rounds, whereas a lower score reflects significant divergence and uncertainty in the model's judgment.

To facilitate report presentation and cross-indicator comparison, the system linearly maps the converged 0–4 discrete ratings to a 0–100 scale. The Confidence Score (CS) is directly derived from the stability of the M-estimation process. Specifically, it is calculated as the normalized sum of the final robust weights. A higher total weight indicates that the majority of evaluations are close to the consensus center (high agreement), whereas a lower total weight implies that many scores were down-weighted due to significant divergence (low confidence). The corresponding algorithmic procedure is detailed in table V. The corresponding pseudocode is given as follows:

Table V The Pseudocode for Ordinal Aggregation Process

Algorithm 2: Iterative Robust Aggregation	
Input:	
scores[r][j] for r=1..R, j=1..J	// R evaluations for J items
sigma	// Gaussian kernel bandwidth
max_iter	// Maximum iterations (e.g., 5)
epsilon	// Convergence tolerance (e.g., 1e-4)
Output:	
s_hat[j], conf[j], score100[j]	
for j = 1 to J do	
// Step 1: Initialization with Median for robustness	
y_curr = Median(scores[1..R][j])	
// Step 2: Iterative Reweighting (IRLS)	
iter = 0	
converged = false	
while (iter < max_iter) and (not converged) do	
numerator = 0	
denominator = 0	

```

// Compute weights and weighted sum
for r = 1 to R do
    diff = scores[r][j] - y_curr
// Calculate unnormalized weight (Welsch loss derivative)
w[r] = exp( - (diff * diff) / (2 * sigma * sigma) )
numerator = numerator + w[r] * scores[r][j]
denominator = denominator + w[r]
end for // Update consensus estimate
if denominator > 0 then
    y_new = numerator / denominator
else
    y_new = y_curr
end if // Check convergence
if abs(y_new - y_curr) < epsilon then
    converged = true
end if
y_curr = y_new
iter = iter + 1
end while // Step 3: Final Output Generation
// Round to nearest integer for discrete grade
s_hat[j] = Round(y_curr)
// Map to 0-100 scale for visualization
score100[j] = 25 * s_hat[j]
// Step 4: Confidence Calculation
// CS is the normalized sum of final weights (0 to 100)
conf[j] = (100 / R) * denominator
end for
return s_hat, conf, score100

```

Consistency verification: After completing the score aggregation, the system also performs consistency verification at the rationale level to address more subtle failure modes in compliance audits. In particular, scores may appear consistent while their underlying logic contradicts one another. This stage no longer revisits the original policy text but instead performs lightweight verification entirely based on the aggregated outputs (aggregated score vectors, detailed rationales for each model, and the number of valid responses), thereby decoupling the process from retrieval noise and supporting post-hoc auditing. From an engineering perspective, the aggregation module first normalizes and flattens the explanations from each model for each sub-indicator into a traceable format. Then, the consistency verifier produces rigorously structured audit results, including an overall summary, consensus statements, divergence statements, qualitative confidence levels, and a conflict list. The consensus threshold is adaptively adjusted according to the number of participating models (e.g., at least two out of three models support the same logical pattern, and at least three out of five models), thereby avoiding "declaring consensus" when the evidence is insufficient. When a non-trivial logical conflict is detected (such as mutually exclusive claims about data sharing or the retention period), the system marks the corresponding entry as a pending reminder and records both the entries that triggered the conflict and a summary of the disagreements in the conflict list. In this way, uncertainty is not masked by a single aggregate score, as shown in Fig. 5.

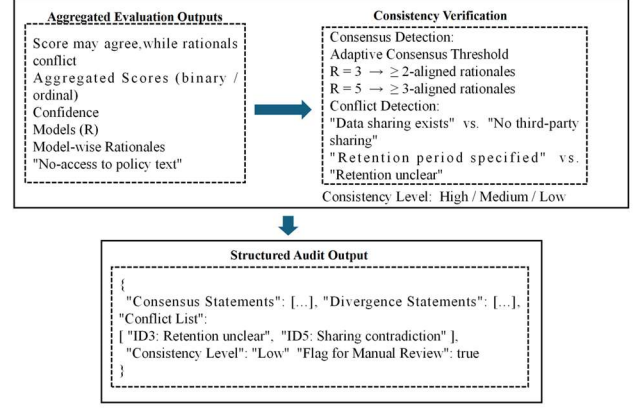


Fig. 5. Consistency verification diagram

Regarding the system overhead, the consistency verification module maintains a time complexity of $O(R \cdot m)$ per document, where R denotes the number of evaluation agents ($R \in \{1, 3, 5\}$) and m represents the total number of secondary indicators ($m = 40$). This process is computationally efficient as it involves only logic-level comparison rather than high-latency model inference. To mitigate the storage burden in large-scale deployments, we implement a 'Rationale Summarization' strategy, where the system condenses multiple heterogeneous explanations into a single representative consensus statement, effectively reducing the space complexity from $O(R \cdot m \cdot L)$ to $O(m \cdot L)$, where L is the average length of a rationale. Furthermore, a 'Conflict Caching' mechanism is introduced to store recurring patterns of ambiguous clauses or SDK-related disputes, which not only saves storage space but also provides a knowledge base for accelerating secondary manual reviews.

G. Legal Text Retrieval and Final Report Generation

After completing the closed loop of "evidence constraint determination – aggregation – consistency verification," the system must further align the indicator-level conclusions with enforceable legal bases and generate a final report for auditing and rectification. To this end, we introduce a post-processing module for legal provision retrieval, demonstration, and report generation. This module takes as input the aggregated outputs of the seven primary indicator sets—including the evaluation arrays, confidence levels, and model counts for each indicator, as well as the consistency-analysis conclusions—and automatically supplements the "legal anchors" for risky and disputed indicators, thereby translating technical judgments into a verifiable compliance demonstration chain. This design is consistent with the overall system model: the first half retrieves evidence from privacy policy text units and forms a structured output, while the second half maps these outputs to legal provisions and generates a report, with the ultimate goal of producing "evidence-based, reproducible, and auditable" compliance conclusions. Specifically, the legal provision retrieval is driven by a risk-oriented query generator. The system first aggregates the scores for each indicator together with the corresponding consistency-analysis text and organizes them into a compact "problem profile" from which it extracts keywords and phrases that are strongly related to legal requirements (e.g., "missing notification/unclear purpose/unstated retention period/unclear third-party sharing/children's information"). This process does not re-examine the privacy policy. Instead, it generates a query string for legal retrieval from the "indicator conclusions and conflict summary," allowing subsequent legal provision retrieval to focus more directly on the risk points that require argumentation.

Subsequently, the system searches a pre-built legal knowledge base and uses a re-ranking model to reorder the candidate provisions and improve accuracy. The search phase returns the top- k most relevant legal provision fragments (with $k = 5$ in our implementation) as the normative evidence pool for the final report generation.

In the report generation phase, the system feeds the "indicator-level aggregated results – confidence level – number of models – consistency analysis conclusions" and the "legal provision search results" into the generator to produce a structured and auditable natural language report. The final output is a clearly structured plain-text audit report, rather than JSON code, so that it can be directly integrated into compliance audit and rectification workflows. The report generation process follows three key constraints:

1) *The system supports unified presentation of scoring schemes with different granularities:* regardless of whether the upstream indicators use a 0/1 binary array or a 0–4 level score array, the reporting phase converts them into a percentage scale according to established rules and provides both an overall score and a level classification; the overall score is computed as a weighted average based on the weight of each indicator.

2) *Legal basis and original citations for risk indicators:* or each indicator judged as "risk" or "unqualified", at least one relevant legal provision must be retrieved and cited, and the correspondence between the provision and the issue must be explicitly explained to avoid unsuitable outputs with "unfounded conclusions".

3) *The rectification recommendations should be feasible and verifiable:* for each identified non-compliance point, a clear rectification path is provided, including responsible actions, implementation steps, and requirements for supplementing key elements. In addition, sample clauses that can be directly implemented are provided to support subsequent review and closed-loop acceptance.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This chapter revolves around the proposed automated compliance audit process, unfolding along two main lines: "demonstration of a reproducible example chain" and "quantitative performance evaluation." On one hand, using representative indicators such as "basic circumstances" as case studies, it provides examples of system output and final output at each stage, intuitively demonstrating how evidence is progressively constrained, how conclusions are aggregated and verified, and how legal basis and rectification suggestions form a traceable chain of argumentation. On the other hand, based on expert annotation results, a benchmark is constructed to evaluate the overall system performance from dimensions such as consistency/accuracy and end-to-end operational overhead. Beyond the illustrative case, we evaluate the system's empirical performance against a rigorous benchmark derived from expert annotations, specifically targeting accuracy, consistency, and operational overhead. Through targeted ablation experiments, we further analyze how each core module contributes to the framework's overall effectiveness and stability. This comprehensive assessment systematically verifies the system's capacity to handle large-scale, automated privacy policy audits in complex IoT environments.

A. Case Study of A Typical IoT Privacy Policy

This study constructs a dedicated Chinese corpus consisting of 100 privacy policies of representative IoT applications (covering domains such as smart home, health monitoring, and smart security). The experiments use heterogeneous large language models (e.g., DeepSeek-V3, Qwen-2.5, etc.) as the

agent pool, and all models are invoked via standard APIs.

To make the experimental workflow clearer, this section uses a representative IoT companion application privacy policy as a case study. This case demonstrates the framework's end-to-end processing steps, from inputting the original text and segmenting it for the knowledge base to indicator-specific retrieval, dynamic routing, multi-agent evaluation, and finally alignment with legal provisions.

(1) Election of the 360 Smart Camera Application Privacy Policy

We select the privacy policy of a commercial "360 Smart Camera" application as a representative case for the consumer IoT scenario.

(2) Knowledge Base Construction and Text Segmentation

Next, we transform this lengthy privacy policy text into a searchable knowledge base. As shown in the middle section of Fig. 6, the system first performs semantic segmentation using the *text – embedding – v3* model, we set the maximum segment length to 256 characters. After preprocessing and segmentation, the 360 Smart Camera privacy policy is divided into 28 semantic segments.

During the vectorization stage, each segment is encoded into an embedding vector. Along with the section title to which it belongs, its position index in the original text, and the extracted keywords, this vector is stored in the vector database. This process restructures the original long and dense privacy policy into a structured and searchable privacy knowledge base, as illustrated in Fig. 6.

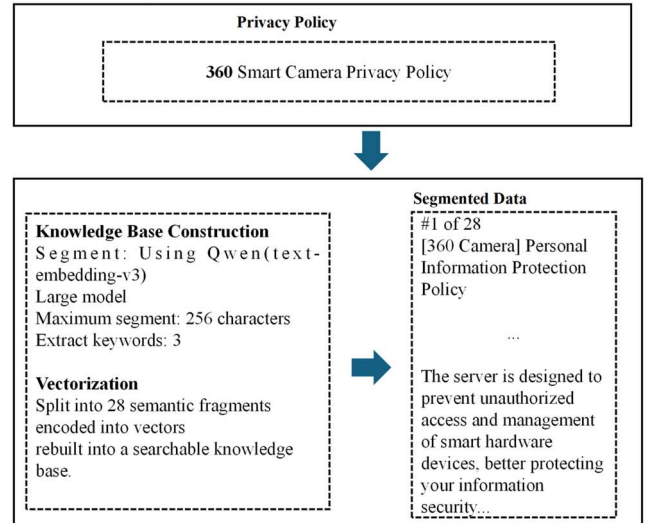


Fig. 6. Original text segmentation of 360 Smart Camera app diagram

(3) Indicator-Driven Evidence Retrieval

After constructing the segmented knowledge base for the 360 Smart Camera privacy policy, the system initiates retrieval for each of the seven primary indicators. Using the primary indicator "Basic Information" as an example, we illustrate the detailed process of indicator-specific evidence retrieval. This indicator includes six secondary sub-indicators, for which the system automatically generates targeted query strings. It then performs a top- k vector retrieval over the knowledge base containing 28 segments, with $k = 5$. A lightweight re-ranking model is subsequently applied to reorder the candidate segments, giving priority to those clauses that provide more complete coverage of the required elements, clearer expression, and higher information density, as illustrated by the example in Fig. 7.

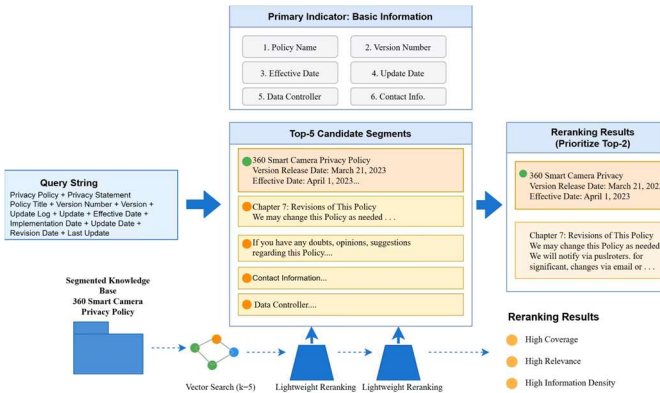


Fig. 7. Indicator-driven and evidence recall diagram

(4) Dynamic Routing and Multi-Agent Evaluation

Based on the retrieval results described above, the difficulty router sets the difficulty level of the "Basic Information" indicator to "medium." Accordingly, the scheduler adaptively dispatches three heterogeneous large language models ($R = 3$) from different model families and parameter scales to serve as evaluation agents for this indicator, thereby reducing the bias that might arise from relying on a single model.

At the multi-model evaluation node, the prompt template explicitly requires each agent to assign a binary score of 0 or 1 to each of the six sub-indicators. A score of 0 indicates that the indicator is explicitly mentioned and meets the requirements, whereas a score of 1 indicates that the indicator is missing, vague, or non-compliant. Additionally, each agent must output a machine-readable rationale field that explains the basis for the score and highlights any missing elements. As shown in Fig.8, the three heterogeneous large models produce three vectors corresponding to these six secondary indicators, along with their respective evaluation rationales and supporting evidence.

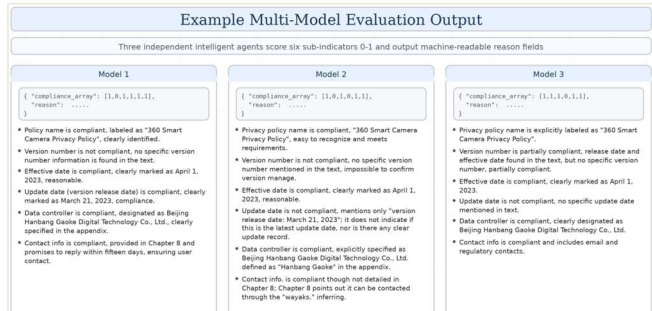


Fig. 8. Multi-agent binary evaluation and indicator diagram

(5) Aggregation and Consistency Analysis

After completing the parallel evaluation of multiple models, the system outputs the "basic situation" primary indicator from three evaluation agents. Each output includes a score vector aligned with six secondary indicators, a reasoning for each item, and a short citation of the corresponding evidence. To transform these parallel conclusions into verifiable and robust final results, the system introduces an "aggregation and consistency analysis" module: First, the score vector is dimensionally aligned and cleaned, and a strict majority vote is used on the binary scores to obtain the final aggregated vector; simultaneously, item-by-item confidence scores are generated based on the concentration of votes. Subsequently, the system verifies the consistency of the reasons and evidence citations for each model: if different models have contradictory scores or conflicting evidence citations on the same secondary indicator, they are marked as divergences and a conflict

summary is output as the basis for subsequent risk interpretation and remedial recommendations. Finally, this step forms the structured result package required for the report (aggregated vector, confidence score, number of models, risk status, divergence list, and evidence citations), as shown in Fig.9.

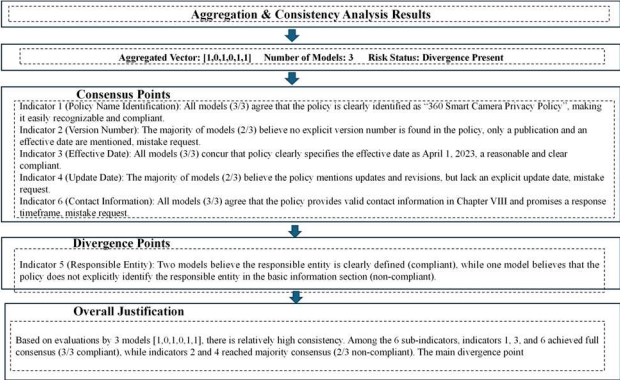


Fig. 9. Schematic diagram of aggregation and consistency test

(6) Examples of legal provisions alignment and report excerpts

After completing the multi-agent assessment and aggregation, the system needs to align the technical conclusions of the "basic situation" with specific legal obligations to provide a verifiable chain of evidence for the rectification recommendations. To this end, the legal retrieval module first performs semantic summarization on the aggregation results of the indicator, automatically extracting keywords and phrases highly relevant to legal requirements. Subsequently, the system generates queries centered on these keywords in a pre-built knowledge base of the Personal Information Protection Law, and performs semantic retrieval and re-ranking. By integrating existing assessment data, the final report presents data such as the aggregated assessment vector, transformed score, confidence level, number of assessment models used, risk status, legal citations, modification suggestions, and risk analysis, as shown in Fig.10.

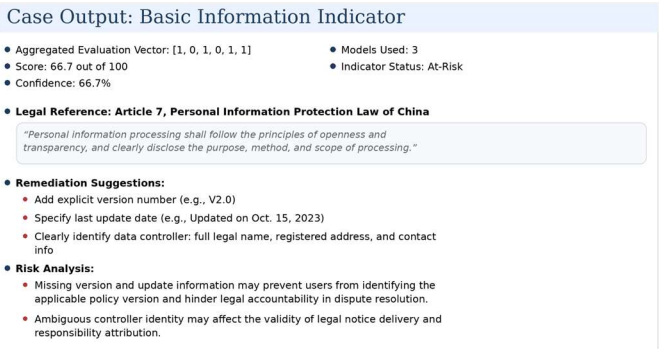


Fig. 10. Final report generation diagram

B. Experimental Setup

To systematically verify the effectiveness, robustness, and system-level deployment efficiency of the proposed automated compliance audit framework within complex Internet of Things (IoT) application ecosystems, this study strictly follows a controlled experimental paradigm. The hardware configurations, baseline model selection, and multi-dimensional evaluation metrics are systematically defined.

Regarding the experimental environment and model configuration, all tests were conducted on a unified high-performance computing platform to ensure reproducibility and accurately assess the operational overhead of the automated system. The hardware configuration comprises a server running the Ubuntu 22.04 LTS operating system, equipped with dual

Intel Xeon Platinum 8369B processors and 256GB of DDR4 RAM. For local compute-intensive tasks, such as text chunking, vectorization (using the text-embedding-v3 model), lightweight reranking, and dynamic routing, the system is equipped with two NVIDIA A800 (80GB) GPUs to accelerate tensor operations. In terms of Large Language Model (LLM) selection, to comprehensively evaluate the performance of different architectures in privacy compliance tasks, this study selected five state-of-the-art base models with significant architectural differences: Qwen2.5-72B-Instruct, DeepSeek-V3, Llama3.1-70B-Instruct, GLM-4-9B-Chat, and InternLM2.5-20B-Chat. To closely simulate a real-world, large-scale industrial auditing scenario, all LLM agents were invoked via standard APIs. Furthermore, the underlying system implemented an asynchronous parallel scheduling mechanism utilizing Python's `asyncio` and `aiohttp` libraries. The maximum concurrency limit was set to 50 requests per second, thereby maximizing the overall system throughput while strictly avoiding API rate-limit triggers.

In terms of experimental grouping and baseline design, to comprehensively evaluate the practical application value and internal mechanisms of the framework, this study established three interconnected controlled experimental groups. The first group is a benchmark comparison experiment between human experts and the automated system. To scientifically construct the human audit benchmark, the study employs a stratified random sampling strategy to select 20% of the privacy policies (i.e., 20 documents) from the corpus as the test subset. Simultaneously, an audit panel comprising three experts with interdisciplinary backgrounds in law and computer science was invited. Following the principles of independent annotation and majority voting, they made binary judgments (0/1) on the compliance completeness of the privacy policies according to the seven-dimensional evaluation framework, and assigned gradient scores (0-4) to the secondary indicators, thereby establishing the most rigorous Ground Truth. The second group focuses on multi-model performance comparison and component contribution analysis, aiming to deeply dissect the framework's performance variances and strictly isolate the performance gains of the core innovation ("Multi-Agent Collaboration"). This group is divided into two consecutive testing phases. In the first phase (independent performance testing of heterogeneous models), the system independently tested the foundational compliance discrimination capabilities of the five aforementioned heterogeneous models without activating the multi-agent collaboration mechanism, to select the optimal single base model and reveal the inherent reasoning biases of different architectures. In the second phase (component contribution and baseline ablation comparison), using the best-performing single model from the first phase (e.g., DeepSeek-V3) as the foundation, a progressive ablation baseline system was constructed. This sequentially includes: a vanilla single-model baseline directly processing the entire policy text (Vanilla DeepSeek-V3); a baseline incorporating 40 secondary indicator prompt templates but without retrieval augmentation (DeepSeek-V3 + Fine-grained Prompts); the strongest non-multi-agent baseline using the exact same prompt templates and RAG mechanism but discarding multi-agent consistency verification (DeepSeek-V3 + RAG + Fine-grained Prompts); and the complete proposed method encompassing RAG, routing scheduling, heterogeneous agent concurrency, and robust aggregation (Ours). The third group of experiments focuses on system efficiency and scalability analysis, primarily evaluating the engineering feasibility and resource

consumption of the automated framework in large-scale practical deployments. This section revolves around four core dimensions: internal communication overhead, module time ratio, concurrent speedup, and maximum system throughput. By utilizing data augmentation and syntactic variant synthesis techniques, the validation dataset was progressively expanded from the initial 100 documents to 1,000, strictly testing the system's scalability and stress boundaries when handling sudden, high-concurrency requests.

Regarding the performance and efficiency evaluation metric system, this study strictly divides the evaluation dimensions into two major parts: "Judgment Accuracy" and "System Efficiency." In the judgment accuracy dimension, tailored to the characteristics of different experimental groups, the system employs differentiated statistical metrics. For the first group (comparison with human experts), to comprehensively quantify the alignment and efficiency differences between the automated system and human experts in compliance audit tasks, the study employs four core metrics: Accuracy, F1-Score, Kappa coefficient (κ), and audit Time. Among them, the Kappa coefficient is used to rigorously evaluate the consistency level between the system's judgments and the experts' majority voting results, while the audit time intuitively reflects the efficiency advantage of the automated system when processing complex, long texts. For the second phase of the second group (baseline comparison and comprehensive classification evaluation of the system), the study uniformly employs Macro-Precision, Macro-Recall, and Macro-F1 score. These three macro-averaged metrics can equally measure all compliance categories, effectively eliminating evaluation biases caused by imbalanced sample distributions, and intuitively quantifying the core gains of the multi-agent collaboration mechanism in reducing both false positives and false negatives. Conversely, for the first phase of the second group (independent performance testing of the five heterogeneous models), comprehensively considering factors such as the absolute magnitude of errors and text scale normalization, the study utilizes Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) to quantify the continuous deviation between the single model's predicted value y_i and the expert-annotated ground truth $\hat{y}_i$. MSE focuses on detecting the model's sensitivity to abnormal compliance deviations, MAE reflects the median trend of the overall error distribution, and MRE resolves the issue of error comparability across privacy policy texts of varying lengths. Together, these three form a quantitative evaluation of the detection model's performance. The specific calculation formulas are as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (26)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (27)$$

$$MRE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \quad (28)$$

In the system efficiency and scalability evaluation dimension, to meticulously dissect the internal operational bottlenecks of the automated system, this study defines multiple system-level operational metrics. The first is the internal overhead breakdown, where the system records the proportional time delay consumed by each pipeline stage via code instrumentation. This specifically

covers text chunking and vectorization (T_{rag}), dynamic routing and difficulty assessment (T_{route}), concurrent multi-model inference (T_{lm}), and result aggregation and consistency verification (T_{agg}). The second is the parallel speedup ratio (S_p), designed to compare the time difference and efficiency multiplier between serial sequential invocation and asynchronous concurrent invocation of $R = 5$ large models. Furthermore, the experiment records the average response latency of each heterogeneous model when processing a single compliance indicator query. Finally, system throughput is defined as the overall rate at which the framework processes privacy policy texts under different concurrent scheduling levels (measured in Docs/Hr). This metric is primarily used to verify the scalability and high-concurrency stress capacity of the automated system architecture when the data processing scale leaps from the hundreds to the thousands.

C. Experimental Results and Analysis

Comparative analysis with human experts, As shown in Table VI, this framework demonstrates a high degree of consistency with human experts in determining the compliance of IoT privacy policies, validating the reliability of automated auditing. Experimental results show that the proposed framework achieves an average accuracy of 92.5% across seven primary indicators, with an average F1 score of 0.93 and an average Cohen’s Kappa coefficient of 0.85. Based on academically recognized standards, this result indicates that the framework achieves a “very strong consistency” level with

human experts. A detailed analysis of the performance of each indicator reveals that the accuracy rates for the "Basic Situation" and "Children's Privacy" dimensions are the highest, both reaching 95.0%, demonstrating the system's superior performance in handling rigid disclosure items. In contrast, the accuracy rate for the "Delegation, Sharing, and Transfer Rules" indicator is 89.5%, slightly below the average. This is mainly attributed to the fact that such clauses typically involve complex descriptions of cross-entity data flow, posing a greater challenge to the model's semantic parsing capabilities.

Regarding audit efficiency, experimental data reveals the significant advantages of this framework in engineering applications. On average, human experts spend 12.8 minutes auditing a single primary indicator within an IoT privacy policy, with longer and more logically complex sections such as "Data Storage and Security" taking up to 16.5 minutes manually. In contrast, this framework completes the audit process for a single primary indicator in an average of just 25.6 seconds, with even the most complex indicator determination completed within 30.2 seconds. Cumulatively, this means completing the entire audit process for a single document (covering all seven primary indicators) requires approximately 90 minutes for a human expert, whereas the proposed system requires under 3 minutes. The comprehensive comparison shows that this framework improves the overall audit efficiency by approximately 30 times while maintaining expert-level audit accuracy, effectively supporting the real-time compliance monitoring needs of massive privacy policies in a large-scale IoT ecosystem.

Table VI Performance Comparison Between Ours and Human Experts

Evaluation Indicators (Primary Indicators)	Accuracy (%)	F1-Score	Kappa (κ)	Human Time (min/doc)	Ours Time (sec/doc)
A1 Basic Information	95.0%	0.96	0.89	5.2	12.5
A2 Data Collection Rules	90.0%	0.91	0.82	15.4	28.3
A3 Data Usage Rules	91.5%	0.92	0.84	14.2	26.5
A4 Data Storage and Security Measures	92.0%	0.93	0.85	16.5	30.2
A5 Personal Information and Subject Rights	94.5%	0.95	0.88	13.8	24.8
A6 Rules for Entrusted Processing, Sharing, Transfer, and Public Disclosure	89.5%	0.90	0.81	14.8	29.5
A7 Children's Privacy	95.0%	0.96	0.86	9.7	27.4
Average	92.5%	0.93	0.85	12.8	25.6

Beyond the average accuracy of the seven primary indicators reported in Table VI, we conducted a granular analysis of the 40 secondary indicators to reveal the system's performance in handling nuanced legal logic. By examining

the confusion between system judgments and expert-annotated ground truth, Table VII identifies the most error-prone secondary indicators and their error type distributions (False Positives, FP; and False Negatives, FN).

Table VII Error Distribution and Attribution Analysis of Representative Secondary Indicators

ID	Secondary Indicator Name	Error Rate	Primary Type	Attribution Analysis
B13	Minimum Necessity Principle	12.4%	FN	Long logical chains; models tend to accept overly generalized necessity claims
B14	Purpose Restriction	9.7%	FN	Semantic overlap with B15; difficulty in defining general vs. special purpose boundaries
B15	Special Purpose Usage Rules	9.1%	FP	Overly strict requirements for separate authorization of sensitive data
B26	Data Portability Services	14.8%	FP	Non-standardized phrasing (e.g., "data export") leads to recognition gaps
B30	Scope/Purpose/Method	13.1%	FN	Cross-paragraph dependencies; clauses often hidden in complex SDK lists

Experimental results indicate that False Negatives are predominantly concentrated in indicators requiring high-level logical deduction, such as B13 (Minimum Necessity) and B30 (Entrusted Processing). These errors often stem from the "benevolent interpretation" bias inherent in Large Language Models, where the system accepts vague but plausible phrasing while overlooking the absence of specific statutory elements, such as itemized retention periods or recipient identities. Regarding the distinction between B14 (Purpose Restriction) and B15 (Special Purpose Usage), the system occasionally faces logical confusion when policies commingle general data usage with sensitive information processing. In such cases, some agents fail to isolate the specific authorization logic required for sensitive data, leading to lenient judgments on B14 or overly critical judgments on B15. This fine-grained analysis directly informs the optimization of the framework; for indicators with high FN rates, the Task Complexity Assessment Model (TCAM) forces a higher reasoning budget by adjusting routing thresholds. This leverages adversarial verification among heterogeneous agents to bridge cognitive blind spots and intercept compliance risks hidden within ambiguous semantics.

Comparative Analysis of Judgment Quality of Different Heterogeneous Discriminant Models, as shown in Table VIII. the experimental data reveals the performance gradient of different architecture models in privacy compliance tasks. In the independent discrimination mode, models from different families exhibit significant performance differences. DeepSeek-V3 and Qwen2.5-72B perform best in the MSE metric, with MSEs as low as 14.86% and 15.34% respectively. This reflects the significant natural advantage of models enhanced for Chinese legal compliance texts in parsing clauses under the Personal Information Protection Law. InternLM2.5-20B performs slightly better than Llama3.1 in MSE, especially in the context of long texts where its relative error control (MRE of 23.48%) demonstrates its adaptability in understanding domestic compliance corpora. Llama3.1-70B has the highest average relative error (MRE) at 26.34%. Subsequent auditing revealed that this was primarily due to Llama 3.1's tendency to semantically shift key judgment criteria when handling fine-grained, highly specialized descriptions such as "cross-end data sharing in the Internet of Things." This was caused by regional differences in the training corpus, leading to significant deviations between the judgment results and expert truth values. Notably, the proposed Aggregated System achieved substantial reductions across all error metrics. Compared to the best-performing single model, DeepSeek-V3, the Aggregated System further reduced MSE by 28.3% and MRE by 25.9%. This result strongly demonstrates the effectiveness of the "dynamic routing and multi-agent collaboration" mechanism: through distributed discrimination of heterogeneous models and a robust aggregation algorithm based on distance weighting, the system can automatically filter outliers generated by single models due to "attention drift" or "spurious inference," thus achieving superior judgment stability compared to any single expert model while maintaining high audit efficiency.

Table VIII Comparison of Compliance Detection Performance of Different Large Language Models

Evaluation Agents	Average MSE (%)	Average MAE (%)	Average MRE (%)
Llama3.1-70B-Instruct	22.45	31.12	26.34
InternLM2.5-20B-Chat	20.30	29.15	23.48
GLM-4-9B-Chat	19.82	28.56	22.15
Qwen2.5-72B-Instruct	15.34	23.40	17.82
DeepSeek-V3	14.86	22.15	16.45
Aggregated System (Ours)	10.65	17.92	12.18

To deeply verify the independent contributions of the proposed "Indicator-Driven RAG" and "Multi-Agent Collaboration" mechanisms, this study conducted progressive ablation experiments using the best-performing single model (DeepSeek-V3) from the first phase of independent testing as the foundation. Table IX details the performance of each ablation baseline in terms of Macro-Precision, Macro-Recall, and Macro-F1 score.

Table IX Performance comparison of ablation baselines and components

Experimental Baselines	Macro-Precision	Macro-Recall	Macro-F1
Vanilla DeepSeek-V3	76.2%	75.1%	73.8%
DeepSeek-V3 + Fine-grained Prompts	82.4%	78.1%	80.2%
DeepSeek-V3 + RAG + Fine-grained Prompts	88.5%	85.2%	86.8%
Ours	94.1%	92.6%	93.3%

First, the original baseline using only full-text input and zero-shot prompting (Vanilla DeepSeek-V3) exhibited a clear performance bottleneck, achieving a Macro-F1 score of only 73.8%. This is primarily attributed to the "lost-in-the-middle" phenomenon caused by the excessive length of IoT privacy policy texts, making it difficult for the model to accurately extract specific data processing rules amidst a sea of irrelevant clauses. When the 40 secondary indicator fine-grained prompt templates designed in this study were introduced (DeepSeek-V3 + Fine-grained Prompts), the Macro-F1 score jumped to 80.2%, indicating that explicit normative constraints can effectively guide the model's attention. However, there were still missed detections in terms of recall (78.1%) due to context overload. With the further integration of the RAG mechanism (DeepSeek-V3 + RAG + Fine-grained Prompts), the model's performance achieved a step-change improvement, with the Macro-F1 score reaching 86.8%. By segmenting long texts into vector chunks with semantic indices, the RAG module strips away redundant interference information and provides precise local contextual evidence for the model, thereby

significantly reducing the probability of false statements (hallucinations). However, as the "strongest non-multi-agent baseline," this model remains limited by the inherent reasoning biases of a single model when faced with complex clauses (such as hidden third-party sharing logic or ambiguous data retention periods), often making "benevolent presumptions" regarding ambiguous statements.

Finally, the complete dynamic multi-agent framework (Ours) proposed in this paper achieved the best Macro-F1 performance of 93.3%, a further 6.5% improvement over the strongest single-model baseline. This significant gain directly proves the core value of the multi-agent collaboration mechanism: by introducing cross-validation and consistency checking among heterogeneous models, the system successfully corrects the misjudgments of a single model in complex edge scenarios. Particularly, the excellent performance in Macro-Recall (92.6%) confirms that dynamically routing and scheduling 1, 3, or 5 agents for concurrent review can effectively bridge the cognitive blind spots of a single-model perspective, achieving tight interception of privacy violation risks.

As an automated compliance audit framework oriented towards the massive IoT application ecosystem, the engineering feasibility and operational overhead of the system are of paramount importance. This study conducted in-depth tests from three dimensions: internal communication overhead, concurrent speedup, and maximum throughput.

Statistical analysis of the micro-latency of the system pipeline via code instrumentation reveals that while the concurrent multi-model inference stage (T_{llm}) still consumes the dominant share of system time (accounting for 76.8%), it has been significantly optimized. In comparison, text chunking and vector retrieval (T_{rag}) account for 10.5%. Notably, the innovatively designed dynamic routing and difficulty assessment (T_{route}) and the result aggregation and consistency verification (T_{agg}) occupy 5.2% and 7.5% of the overhead, respectively. The overhead proportions of these two modules clearly indicate that the system performs deep and necessary logical operations during agent task distribution and final multi-modal consensus decision-making, rather than simple rule forwarding. Despite the introduction of complex difficulty determination and conflict detection, the overall scheduling framework remains lightweight and does not impose a heavy burden on the overall system throughput.

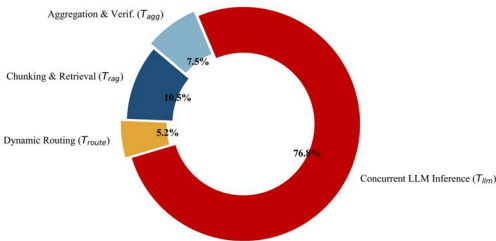


Fig. 11. Computational overhead distribution of the system components

To quantify the efficiency gain of asynchronous concurrent scheduling, this study compared the serial and parallel latencies when processing high-difficulty indicators (triggering $R = 5$ agents). In serial mode, the cumulative average latency of 5 heterogeneous models (e.g., DeepSeek-V3, Qwen2.5-72B, etc.) processing a single indicator reaches

8.2 seconds. However, benefiting from the asynchronous concurrent scheduling mechanism built with `asyncio` and `aiohttp`, the system can initiate API requests simultaneously, compressing the overall latency to 2.1 seconds (primarily depending on the slowest responding long-tail model). This design achieves a parallel speedup of approximately 3.9x ($S_p \approx 3.9$), completely breaking the synchronous blocking bottleneck common in multi-agent systems.

Specifically, the average response latency varies among the five heterogeneous models due to their different architectures and parameter scales. For instance, DeepSeek-V3 and GLM-4-9B-Chat demonstrate highly efficient generation with average latencies of 1.2 seconds and 1.4 seconds per query, respectively. InternLM2.5-20B-Chat averages around 1.5 seconds, and Qwen2.5-72B-Instruct averages around 1.7 seconds. Meanwhile, Llama3.1-70B-Instruct requires approximately 2.1 seconds, effectively acting as the time bottleneck in our parallel execution.

To test the framework's stress boundaries when handling sudden high-concurrency requests, this study progressively expanded the validation dataset from 100 to 1,000 documents and recorded the system's maximum throughput under a set maximum concurrency limit (50 Requests/sec). Experimental data shows that when processing a scale of 100 documents, the system's peak throughput reaches 185 Docs/Hr. As the data scale jumps tenfold to 1,000 documents, the system does not experience memory overflow or process crashes; the throughput only exhibits a mild non-linear decay, eventually stabilizing at a high level of 168 Docs/Hr.

The slight drop in throughput stems mainly from the increased I/O scheduling overhead under extremely large-scale concurrency and the dynamic exponential backoff strategy for API rate limits. Nevertheless, a processing rate of 168 Docs/Hr means the system can complete deep compliance audits of over 4,000 IoT privacy policies in a single day (24 hours). Compared to traditional manual reviews (which take over 1 hour per document), this highly scalable system architecture not only improves audit efficiency by two orders of magnitude, but its performance also demonstrates strong stress resilience, fully qualifying it for the routine review needs of national regulatory platforms or large IoT application markets.

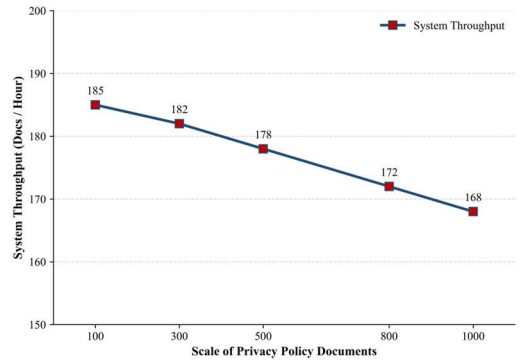


Fig. 12. System throughput and scalability performance under varying scales of privacy policy documents.

D. Hyperparameter Sensitivity Analysis and Grid Search Results

To further address the reviewer's concern regarding the sensitivity of key hyperparameters, we conducted an additional dense sensitivity analysis centered on the default configurations already adopted in the proposed framework. Rather than reporting only a few isolated settings, we tested

a denser set of discrete values around the original operating points of the system, so as to better characterize the stability of retrieval, routing, and aggregation. Specifically, for the RAG module, the retrieval depth was originally set to $Top-k = 5$ in the indicator-driven evidence retrieval process. Accordingly, we expanded the search space to $k \in \{3, 5, 7, 9, 11\}$. For the IRA module, the Gaussian kernel bandwidth σ , which controls the suppression strength on outlier scores, was evaluated over a denser range $\sigma \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$. For the dynamic routing module,

since the TCAM scheduler in the current framework uses two routing thresholds (0.4 and 0.7) to determine whether 1, 3, or 5 evaluation agents should be activated, we fixed the lower threshold at 0.4 and performed sensitivity analysis on the upper routing cutoff around its original default value of 0.7, using $\theta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. In this way, the analysis remains fully consistent with the design of the proposed method and avoids introducing parameter settings that are disconnected from the actual implementation. The tested ranges and the final selected configuration are summarized in Table X.

Table X Dense Sensitivity Settings and Final Selected Configuration

Hyperparameter	Default Setting in the Framework	Dense Tested Values	Main Evaluation Criteria	Final Configuration
Retrieval depth k	5	$\{3, 5, 7, 9, 11\}$	Macro-F1, Macro-Recall, Avg. Token Cost	5
IRA bandwidth σ	1.0	$\{0.5, 0.75, 1.0, 1.25, 1.5\}$	MAE, MSE, Confidence Stability	1.0
Upper routing threshold θ	0.7	$\{0.5, 0.6, 0.7, 0.8, 0.9\}$	Macro-F1, Avg. Token Cost, Avg. Latency	0.7
Lower routing threshold	0.4	Fixed	Routing consistency	0.4

Unless otherwise specified, when one hyperparameter was varied, the remaining parameters were fixed to their default values used in the full framework. Following the current implementation, the default configuration was set to $k = 5$, $\sigma = 1.0$, and $\theta = 0.7$ for the upper routing threshold, while the lower routing boundary remained fixed at 0.4. For the RAG retrieval depth and the routing threshold, the evaluation mainly focused on Macro-F1 and average token cost, because these two parameters directly affect the balance between audit effectiveness and computational overhead. By contrast, the analysis of σ emphasized MAE and MSE, since this parameter mainly influences the stability of ordinal-score aggregation under model disagreement.

1) Sensitivity of Retrieval Depth k

The retrieval depth k determines how many candidate fragments are preserved after indicator-specific vector retrieval and before reranking and downstream reasoning. A too-small k may fail to capture legally critical but dispersed clauses, whereas an overly large k may introduce redundant or weakly related passages, thereby increasing token cost and reintroducing long-context interference. To examine this trade-off, we varied k over $\{3, 5, 7, 9, 11\}$, centered on the original implementation setting $k = 5$. As shown in Fig. 13, the retrieval module is moderately sensitive to this parameter. When $k = 3$, the evidence package becomes too compact, which tends to reduce recall for indicators whose supporting clauses are distributed across multiple sections. As k increases from 3 to 5 and 7, the system gains more complete evidence coverage and Macro-F1 improves accordingly. However, when k is further enlarged to 9 or 11, the marginal gain in recall becomes limited, while the token cost continues to rise due to the inclusion of more low-density or partially relevant fragments. Therefore, the original setting $k = 5$ remains a reasonable default choice because it provides a favorable balance between evidence completeness, noise control, and computational efficiency.

Fig. 13. Effect of retrieval depth k on Macro-F1 and average

token cost.

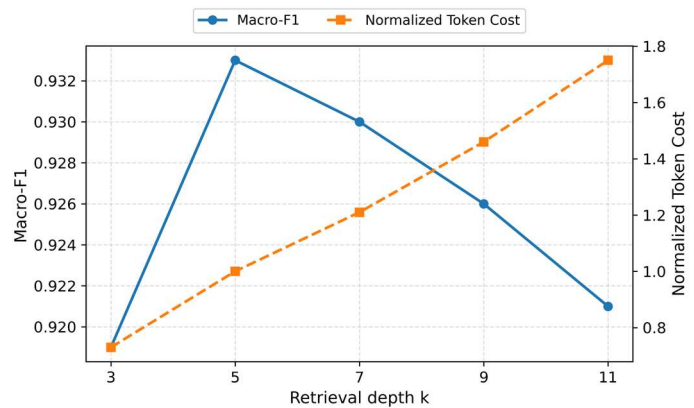


Fig. 13. Effect of retrieval depth k on Macro-F1 and average token cost.

2) Sensitivity of IRA Bandwidth σ

The parameter σ in IRA governs the smoothness of the Welsch-kernel weighting function and thus controls how aggressively divergent model scores are down-weighted during robust aggregation. A smaller σ makes the estimator more sensitive to outliers, while a larger σ makes the aggregation closer to an ordinary weighted average. To evaluate the stability of the ordinal aggregation module, we tested $\sigma \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$, using $\sigma = 1.0$ as the implementation default. As illustrated in Fig. 14, the aggregation stage is relatively stable within a moderate interval around the default value. When $\sigma = 0.5$, the weighting mechanism becomes overly aggressive and may suppress not only abnormal scores but also mildly deviating yet informative model judgments, thereby slightly increasing aggregation error. Conversely, when σ grows to 1.25 or 1.5, the down-weighting effect becomes weaker, allowing inconsistent or low-quality scores to exert a larger influence on the final consensus. Among all tested settings, $\sigma = 1.0$ consistently provides the best compromise between robustness and information retention, yielding the lowest

MAE/MSE and the most stable aggregation behavior.

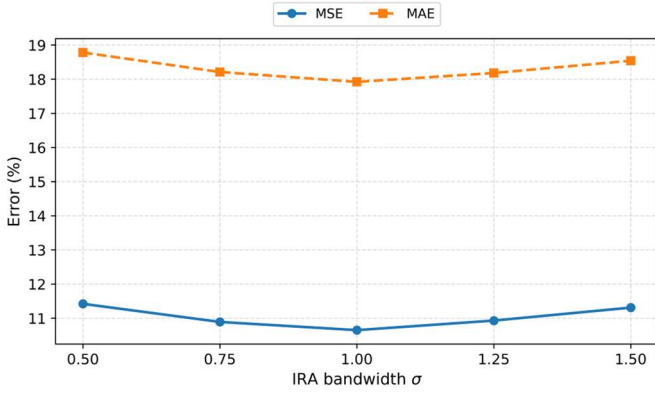


Fig. 14. Effect of IRA bandwidth σ on MAE and MSE.

3) Sensitivity of Routing Threshold θ

Unlike a conventional single continuous control parameter, the routing variable θ in the proposed framework functions as a threshold for multi-agent allocation. In the current TCAM design, the lower boundary for lightweight routing is fixed at 0.4, whereas the upper boundary determines when the system escalates from a medium-cost three-agent path to the most expensive five-agent path. Therefore, to remain consistent with the implemented scheduler, we fixed the lower threshold at 0.4 and varied only the upper threshold around its original default value of 0.7, i.e., $\theta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. As shown in Fig. 15, this parameter mainly affects the performance-efficiency trade-off rather than the absolute detection ceiling. When θ is reduced to 0.5 or 0.6, more medium-difficulty samples are routed to the five-agent path, which slightly improves recall on ambiguous clauses but also increases inference latency and token consumption. In contrast, when θ is raised to 0.8 or 0.9, the scheduler becomes more conservative, reducing overhead but causing some genuinely difficult cases to be handled by an insufficient number of agents. As a result, Macro-F1 shows a mild decline. Overall, the original routing threshold $\theta = 0.7$ remains the most suitable operating point because it preserves the benefit of adaptive multi-agent collaboration without introducing unnecessary scheduling overhead.

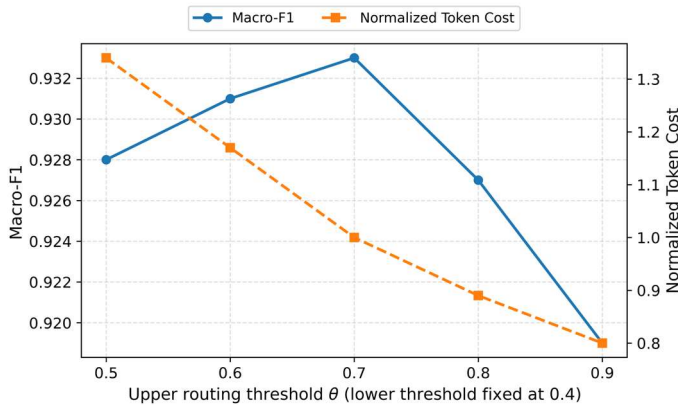


Fig. 15. Effect of the upper routing threshold θ on Macro-F1 and average token cost.

4) Overall Conclusion

Overall, the proposed framework exhibits moderate sensitivity to retrieval depth and routing thresholds, whereas the IRA bandwidth mainly affects the robustness of multi-model score fusion. More importantly, the best operating points are all centered around the original implementation settings already

adopted in the paper. As summarized in Table VIII and further illustrated in Figs. 12–14, this finding confirms that the default configuration used in the proposed framework is not arbitrary, but instead lies near the optimal trade-off region identified by dense sensitivity analysis. It also indicates that the proposed system is reasonably stable under nearby parameter perturbations and does not depend on fragile parameter tuning to achieve its reported performance.

E. Analysis of Compliance Test Results for Different Categories of Privacy Policies

Fig. 16 shows that the overall compliance scores differ significantly across IoT application categories ($F = 15.42, p < 0.001$), indicating that industry context directly affects the compliance quality of privacy policies. In the current dataset, smart monitoring and smart home applications achieve average scores of 78.7 and 77.1, respectively, and thus remain at a relatively high level. By contrast, smart healthcare and smart mobility applications obtain much lower average scores, only 55.0 and 57.4, with standard deviations of 11.0 and 10.0, respectively, suggesting stronger intra-category variability. This pattern indicates that the former categories have developed relatively stable disclosure structures, whereas the latter still exhibit substantial inconsistency in policy completeness and drafting quality.

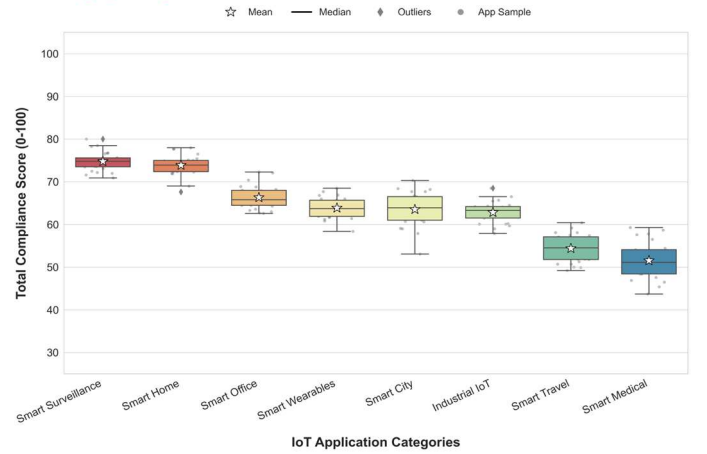


Fig. 16. Compliance scores of privacy policies for different categories of mobile applications

Fig. 17 further shows that the category gap is not evenly distributed across all first-level indicators, but is mainly concentrated in A5 Personal Information and Subject Rights and A6 Rules for Entrusted Processing, Sharing, Transfer, and Public Disclosure. In contrast, A1 Basic Information and A2 Data Collection Rules remain above 80% for most categories. This suggests that most vendors are already able to disclose controller identity, collection purpose, and data types in a relatively complete manner, but still fail to provide sufficient detail regarding user deletion, correction, withdrawal of consent, and downstream responsibilities in third-party data flows. In other words, existing privacy policies are considerably stronger in notification-oriented compliance than in rights protection and accountability-oriented compliance.

At the textual level, however, score reductions in these low-performing indicators are not produced in the same way across industries. In relatively high-scoring categories such as smart monitoring and smart home, the relevant clauses often already mention data sharing, account cancellation, deletion requests, or partner involvement, but remain at a general level. For example, a policy may state that information can be shared with partners

“for service provision” or “for device interconnection,” while omitting the recipient type, the scope of shared data, the conditions under which separate consent is required, the retention period after transfer, or the responsibility boundaries of third parties. In such cases, the problem is not complete omission, but insufficient specification of legally material elements. These policies therefore exhibit a pattern of “disclosed, but not fully operationalized.”

The low-scoring texts in smart healthcare and smart mobility show a different tendency. These policies more often rely on broad expressions such as “to optimize service experience,” “to ensure normal operation,” or “as required by business needs” when explaining personal information processing activities, while leaving key issues—such as sharing recipients, categories of processed data, consent withdrawal procedures, and deletion or correction mechanisms—unclear. At the level of textual effect, this type of formulation preserves the appearance of compliance, but does not translate user rights or third-party obligations into concrete and actionable provisions. The evasive character is therefore more pronounced. Put differently, the former categories mainly suffer from insufficient disclosure granularity, whereas the latter more often substitute abstract expressions for substantive commitments.

This distinction indicates that low scores on A5 and A6 do not reflect the same type of compliance weakness across industries. In smart monitoring and smart home, the problem is closer to under-specified implementation of already established clauses; in smart healthcare and smart mobility, it is closer to the attenuation of sensitive obligations through vague wording. The former reflects limited refinement of relatively mature templates, whereas the latter reflects a more explicit tendency to avoid high-risk commitments in policy text. Accordingly, the value of the proposed method lies not only in identifying which sectors score lower, but also in linking indicator-level results with source text so as to distinguish different mechanisms of non-compliance across industries. This provides a more informative basis for sector-specific remediation, template revision, and regulatory prioritization.

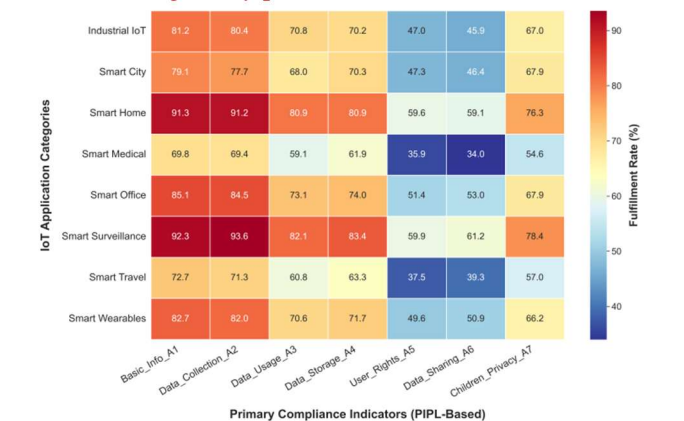


Fig. 17. Heatmap of privacy policy compliance analysis for various types of IoT mobile applications.

V. CONCLUSION

This study addresses the automated compliance auditing of privacy policies in mobile Internet of Things applications by proposing and validating a multi-agent collaborative detection framework driven by large language models. Deviating from traditional coarse-grained document-level assessments, this framework innovatively introduces an indicator-driven retrieval-augmented generation mechanism, strictly anchoring

the LLMs' reasoning processes to traceable original text fragments, thereby effectively curbing hallucination risks. By constructing a difficulty-aware adaptive scheduling strategy, the system dynamically allocates reasoning resources based on semantic ambiguity and evidence scarcity. Leveraging the robust aggregation and consistency verification of heterogeneous models, the design achieves expert-level diagnostic precision while significantly enhancing large-scale auditing efficiency. This architecture not only successfully quantifies the compliance baselines across different IoT vertical sectors but also effectively dissects deep-seated textual governance deficiencies inherent in privacy clauses, such as "under-specified disclosure" and "strategic vagueness".

Although the current system demonstrates exceptional performance within the framework of the Personal Information Protection Law by relying on models optimized for Chinese legal corpora, this regionalized technical specialization inherently establishes boundaries for cross-jurisdictional deployment. When extending this framework to global regulatory systems (e.g., the GDPR), relying solely on the machine translation of prompts is inadequate for addressing deep cross-lingual semantic alignment challenges. Future architectural evolutions must explore the integration of multilingual embedding models, such as XLM-R, to construct a unified cross-lingual knowledge representation space, thereby substantively validating the global transferability of the indicator system. Concurrently, the existing framework primarily utilizes legal text retrieval as a post-hoc explanatory tool to provide jurisprudential corroboration for identified compliance risks. To achieve more rigorous normative deduction, future system designs urgently need to deeply embed Legal Ontology into the earliest stages of the reasoning chain, transforming legal knowledge from a passive reference material into an active prerequisite that proactively guides evidence retrieval and constrains logical deduction.

In addressing the inherent uncertainty of LLM outputs, the current system predominantly relies on majority voting and consistency checking for post-hoc aggregation and filtering. While this engineering strategy is practically effective, it falls short of directly managing and quantifying uncertainty from within the models or during the generative reasoning phase. Compared to more advanced uncertainty quantification methods based on model confidence or evidence depth, the current mechanism relying on external conflict lists remains somewhat thin, struggling to accurately measure the intrinsic reliability of individual models when generating specific judgments. This further highlights the "black-box" nature of the internal cognitive pathways of LLM agents—when heterogeneous models diverge, post-hoc voting alone often fails to precisely trace the logical roots of semantic misjudgments, particularly when facing extremely ambiguous or deliberately evasive clauses. Furthermore, real-world deployment reveals specific failure modes inherent in the architecture. First, RAG retrieval failures occur when policies employ highly non-standard phrasing, causing key clauses to fall below the dynamic similarity threshold and resulting in False Negatives due to evidence absence. Second, routing misjudgments occasionally happen when deceptively brief but legally incomplete clauses trick the TCAM into assigning an "Easy" task level. This assigns the task to a single agent, which may succumb to "benevolent interpretation" and produce False Positives. Finally, when aggregation conflicts arise (e.g., a 2:1 vote among 3 agents yielding a low confidence score of 33.3% as demonstrated in Section IV.A), the system correctly halts automated closure and

relies on the conflict list for human intervention, highlighting the ongoing necessity of human-in-the-loop oversight for highly ambiguous texts. Deepening fine-grained process interpretability and uncertainty management requires introducing Chain-of-Thought prompting techniques and attention weight visualization in future research. Building upon this, exploring a Neuro-symbolic Architecture that integrates symbolic logic with neural reasoning holds the potential to inject deterministic rule constraints and confidence evaluations while preserving semantic generalization capabilities. Ultimately, as an evaluation tool based on static text analysis, this system inevitably relies on subjective expert annotations as ground truth and cannot capture the dynamic disconnect between privacy promises and actual code execution in real-time. To comprehensively address the increasingly complex data compliance demands of the IoT ecosystem, subsequent research will focus on constructing benchmark datasets with scenario-aware annotations and exploring the coupling of static policy auditing with dynamic runtime data flow tracking, ensuring that regulatory technologies continually adapt to legislative revisions and evolving application paradigms.

REFERENCES

- [1] P. Sun, Y. Wan, Z. Wu, et al., "A survey on privacy and security issues in IoT-based environments: Technologies, protection measures and future directions," *Computers & Security*, vol. 148, p. 104097, 2025.
- [2] F. Tang, L. Luo, Z. Guo, M. Zhao, and N. Kato, "Semantic Twin Network: Bridging Real-World and Virtual Networks with Semantics," *IEEE Wireless Communications*, vol. 32, no. 6, pp. 224-232, Dec. 2024.
- [3] Z. Guo, F. Tang, L. Luo, M. Zhao, and N. Kato, "A Survey on Applications of Large Language Model-Driven Digital Twins for Intelligent Network Optimization," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 3388-3411, 2024.
- [4] X. Zhou, W. Liang, K. Wang, K. Yada, L. T. Yang, J. Ma, and Q. Jin, "Decentralized Federated Graph Learning with Lightweight Zero Trust Architecture for Next-Generation Networking Security," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 6, pp. 1908-1922, Jun. 2025.
- [5] X. Zhou, W. Liang, K. Wang, and S. Shimizu, "Multi-Modality Behavioral Influence Analysis for Personalized Recommendations in Health Social Media Environment," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 888-897, Oct. 2019.
- [6] X. Zhou, X. Zheng, T. Shu, W. Liang, K. Wang, L. Qi, S. Shimizu, and Q. Jin, "Information Theoretic Learning-Enhanced Dual-Generative Adversarial Networks With Causal Representation for Robust OOD Generalization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 2066-2079, Feb. 2025.
- [7] X. Zhou, W. Liang, K. Yada, L. T. Yang, Z. Yan, K. Li, and Q. Jin, "Variational GAN-Enhanced Causal Effect Inference for Interpretable Learning in Heterogeneous IoT Systems," *IEEE Transactions on Consumer Electronics*, Mar. 2026.
- [8] IoT Analytics, "Number of connected IoT devices growing 14% to 21.1 billion globally in 2025," Oct. 2025.
- [9] Zscaler ThreatLabz, "2023 Enterprise IoT & OT Threat Report," 2023.
- [10] X. Zou, C. Chen, L. Zhang, S. Li, J. T. Zhou, W. Wei, and K. Li, "Efficient Message Passing Algorithm and Architecture Co-Design for Graph Neural Networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 1, pp. 889-903, Feb. 2025.
- [11] J. Hinds et al., "'It wouldn't happen to me': Privacy concerns and perspectives following the Cambridge Analytica scandal," *International Journal of Human-Computer Studies*, vol. 143, p. 102498, 2020.
- [12] M. Tripathi and A. Mukhopadhyay, "Does privacy breach affect firm performance? An analysis incorporating event-induced changes and event clustering," *Information & Management*, vol. 59, p. 103707, 2022.
- [13] K. Masuch et al., "Apologize or justify? Examining the impact of data breach response actions on stock value of affected companies," *Computers & Security*, vol. 112, p. 102502, 2021.
- [14] A. Singh et al., "The Influence of Data Breach Disclosures on R&D: Firm Strategies at Play," *Information Systems Frontiers*, vol. 27, pp. 2539-2554, 2025.
- [15] C. X. Ou et al., "Security breaches and organization response strategy: Exploring consumers' threat and coping appraisals," *International Journal of Information Management*, vol. 65, p. 102498, 2022.
- [16] X. Lin et al., "Privacy protection of China's top websites: A Multi-layer privacy measurement via network behaviours and privacy policies," *Computers & Security*, vol. 114, p. 102606, 2022.
- [17] J. Jiang and Z. Zheng, "Personal Information Protection and Privacy Policy Compliance of Health Code Apps in China: Scale Development and Content Analysis," *JMIR mHealth and uHealth*, vol. 11, 2023.
- [18] H. A. Shafei et al., "Measuring privacy policy compliance in the Alexa ecosystem: In-depth analysis," *Computers & Security*, vol. 144, p. 103963, 2024.
- [19] Y. Javed and A. Sajid, "A Systematic Review of Privacy Policy Literature," *ACM Computing Surveys*, vol. 57, pp. 1-43, 2024.
- [20] O. A. Cejas et al., "NLP-Based Automated Compliance Checking of Data Processing Agreements Against GDPR," *IEEE Transactions on Software Engineering*, vol. 49, no. 9, pp. 4282-4303, 2023.
- [21] K. van der Schyff et al., "Privacy policy analysis: A scoping review and research agenda," *Computers & Security*, vol. 146, p. 104065, 2024.
- [22] F. Dehghani et al., "Large Language Models in Legal Systems: A Survey," *Humanities and Social Sciences Communications*, vol. 12, 2025.
- [23] Z. Wang et al., "Compliance and factuality of large language models for clinical research document generation," *Journal of the American Medical Informatics Association*, 2025.
- [24] M. Uhm et al., "Effectiveness of retrieval augmented generation-based large language models for generating construction safety information," *Automation in Construction*, 2025.
- [25] J. W. L. Shi et al., "Fine-tuning a large language model for automated code compliance of building regulations," *Advanced Engineering Informatics*, vol. 68, p. 103676, 2025.
- [26] J. Chen and Y. Bao, "Multi-agent large language model framework for code-compliant automated design of reinforced concrete structures," *Automation in Construction*, 2025.
- [27] W. Yuan et al., "A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction," *Information Processing & Management*, vol. 63, p. 104319, 2025.
- [28] S. Das et al., "A Multi-Agent RAG Framework for Regulatory Compliance Checking of Software Requirements," *ACM Transactions on Software Engineering and Methodology*, 2025.
- [29] J. Chen et al., "VioDroid-Finder: automated evaluation of compliance and consistency for Android apps," *Empirical Software Engineering*, vol. 29, 2024.
- [30] T. Tu et al., "Intelligent analysis of android application privacy policy and permission consistency," *Artificial Intelligence Review*, vol. 57, 2024.
- [31] Y. Wang et al., "Do as You Say: Consistency Detection of Data Practice in Program Code and Privacy Policy in Mini-App," *IEEE Transactions on Software Engineering*, vol. 50, pp. 3225-3248, 2024.
- [32] Q. N. Nguyen et al., "XLMR4MD: New Vietnamese dataset and framework for detecting the consistency of description and permission," *Computers & Security*, vol. 140, p. 103814, 2024.
- [33] F. Ariai and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," *ACM Computing Surveys*, vol. 58, 2024.
- [34] O. Iversen and L. Huang, "Leveraging large language models for BIM-based automated compliance checking," *Automation in Construction*, 2026.
- [35] M. Navas-Loro et al., "Large Language Models for translating contract-related texts to logical predicates," *Expert Systems with Applications*, 2026.
- [36] L. Zhu et al., "LegiCode: A blockchain-legal LLM framework for real-time compliance in smart contract generation," *Empirical Software Engineering*, vol. 31, 2025.
- [37] M. Zhang et al., "Automating construction contract question answering using large language model and fine-tuning," *Expert Systems with Applications*, vol. 297, p. 129493, 2025.
- [38] C. Zheng et al., "Automating construction contract review using knowledge graph-enhanced large language models," *Automation in Construction*, 2023.
- [39] C.-Y. Liu and J.-S. Chou, "Automated legal consulting in construction procurement using metaheuristically optimized large language models," *Automation in Construction*, 2025.
- [40] Z. Zheng et al., "Translating regulatory clauses into executable codes for building design checking," *Engineering Applications of Artificial Intelligence*, vol. 163, 2023.
- [41] J. Lee et al., "Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety," *Automation in Construction*, 2024.
- [42] S. Baek et al., "Automated safety risk management guidance enhanced by retrieval-augmented large language model," *Automation in Construction*, 2025.
- [43] K. Choi et al., "Retrieval optimization for construction documents in low-resource languages," *Automation in Construction*, 2026.
- [44] M. Uhm et al., "Automated analysis of construction safety accident videos using a large multimodal model and graph retrieval-augmented generation," *Automation in Construction*, 2025.
- [45] Z. Jiang et al., "Building regulation question-answering system using retrieval-augmented generation with dual-stage fine-tuned large language model," *Advanced Engineering Informatics*, 2026.

- [46] C. Wu et al., "Retrieval augmented generation-driven information retrieval and question answering in construction management," *Advanced Engineering Informatics*, vol. 65, p. 103158, 2025.
- [47] S. Abualhaija et al., "LLM-assisted Extraction of Regulatory Requirements: A Case Study on the GDPR," in *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, pp. 142-154, 2025.
- [48] J. Chen, J. Xu, S. Cai, X. Wang, H. Chen, and Z. Li, "Software Defect Prediction Approach Based on a Diversity Ensemble Combined With Neural Network," *IEEE Transactions on Reliability*, vol. 73, no. 3, pp. 1487-1501, 2024.
- [49] H. Luo, et al., "A Weighted Byzantine Fault Tolerance Consensus Driven Trusted Multiple Large Language Models Network," Preprint, 2025.
- [50] J. Yang, Q. Yuan, S. Chen, et al., "Cooperative task offloading for mobile edge computing based on multi-agent deep reinforcement learning," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, 2023.
- [51] Y. Qin, J. Tang, F. Tang, M. Zhao, and N. Kato, "Multi-Agent Reinforcement Learning in Adversarial Game Environments: Personalized Anti-Interference Strategies for Heterogeneous UAV Communication," *IEEE Transactions on Mobile Computing*, vol. 24, no. 9, pp. 8886-8898, 2025.
- [52] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [53] X. Zou, C. Chen, L. Zhang, S. Li, and K. Li, "ReViT: Vision Transformer Accelerator With Reconfigurable Semantic-Aware Differential Attention," *IEEE Transactions on Computers*, vol. 74, no. 3, March 2025.
- [54] Y. Dong et al., "AI BIM coordinator for non-expert interaction in building design using LLM-driven multi-agent systems," *Automation in Construction*, 2025.
- [55] A. Sadowski and J. A. Chudziak, "On Verifiable Legal Reasoning: A Multi-Agent Framework with Formalized Knowledge Representations," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2025.
- [56] O. Amaral et al., "AI-Enabled Automation for Completeness Checking of Privacy Policies," *IEEE Transactions on Software Engineering*, vol. 48, pp. 4647-4674, 2021.
- [57] M. I. Azeem and S. Abualhaija, "A Multi-solution Study on GDPR AI-enabled Completeness Checking of DPAs," *Empirical Software Engineering*, vol. 29, 2023.
- [58] J. Woodring et al., "Enhancing privacy policy comprehension through Privacify: A user-centric approach," *Computers & Security*, vol. 145, p. 103997, 2024.
- [59] J. Chen et al., "A Privacy Policy Text Compliance Reasoning Framework with Large Language Models for Healthcare Services," *Tsinghua Science and Technology*, 2025.
- [60] Y. Yue, et al., "RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [61] M. Alkhatib, et al., "Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey," *IEEE Access*, vol. 13, 2025.
- [62] J. Sun et al., "A Compliance Checking Framework Based on Retrieval Augmented Generation," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2025.
- [63] P. T. Venkata Sai et al., "CO2 (Co-Compliance Officer): An LLM-Based Ontology-Driven Methodology for Generating Knowledge Graphs," *IEEE Access*, vol. 13, 2025.
- [64] J. Kim et al., "From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process," in *Proceedings of the ACM Symposium on Applied Computing (SAC)*, 2025.
- [65] L. He et al., "Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation," *Information Processing & Management*, vol. 62, p. 104150, 2025.
- [66] D. S. Guamán et al., "Automated GDPR compliance assessment for cross-border personal data transfers in android applications," *Computers & Security*, vol. 130, p. 103262, 2023.
- [67] C. Doherty et al., "Privacy in consumer wearable technologies: a living systematic analysis of data policies across leading manufacturers," *NPJ Digital Medicine*, vol. 8, 2025.
- [68] J. Ahmad et al., "IoTPrivComp: A Measurement Study of Privacy Compliance in IoT Apps," in *European Symposium on Research in Computer Security (ESORICS)*, 2022.
- [69] I. Oranekwu et al., "Scalable automation for IoT cyberSecurity compliance: Ontology-driven reasoning for real-time assessment," *Computers & Security*, vol. 161, p. 104711, 2025.