Technical Report for the OptiGov Project: ARCCS — the LLM-based component

OptiGov Team[1,2]
[1]Instituto de Telecomunicações/IST, Universidade de Lisboa
[2]INESC-ID/IST, Universidade de Lisboa

Abstract

This report introduces **ARCCS**, a multi-agent system for regulation-level compliance assessment. The architecture consists of two coordinated modules, the *Regulatory Processing and Extraction Module (RPEM)* and the *Compliance Classification Module (CCM)*, which jointly support the structured analysis of regulatory texts and proposal documents. The system produces interpretable compliance outputs in the form of regulation-level labels, associated confidence scores, and structured explanations that justify each assessment. ARCCS is designed to provide a scalable and transparent framework for automated compliance analysis across complex regulatory and technical domains. To facilitate practical use, we also provide a lightweight web demo that enables non-technical users to run end-to-end compliance checks via a simple three-step workflow[1].

1 Introduction

Regulatory compliance checking is a critical yet resource-intensive process in domains where technical, contractual, and policy-oriented documents must be continuously evaluated against complex and evolving legal frameworks. The unstructured nature of both regulatory texts and operational documents, combined with the need for fine-grained, explainable decisions, poses significant challenges for scalable automation. Traditional compliance workflows rely heavily on manual legal review, which limits reproducibility, transparency, and adaptability in large-scale or dynamic regulatory environments.

In this technical report, we present **ARCCS** (**A**utomated **R**egulatory **C**ompliance **C**hecking **S**ystem), a *multi-agent pipeline* for regulation-based compliance assessment. As illustrated in Figure 1, ARCCS is designed as a coordinated system that transforms unstructured legal and proposal documents into structured representations and performs regulation-level compliance classification.

The architecture consists of two primary modules: the **RPEM** (**R**egulatory **P**rocessing and **E**xtraction **M**odule) and the **CCM** (**C**ompliance **C**lassification **M**odule). The RPEM is responsible for hierarchically segmenting legal texts into articles, paragraphs, and sub-paragraphs, and for extracting atomic regulatory requirements from each regulatory chunk. This module identifies the legal function, applicable actor, and associated conditions of each provision, and subsequently applies filtering and deduplication strategies to remove semantically overlapping or non-actionable constraints, yielding a refined and operational set of regulatory requirements.

In parallel, the CCM processes the terms-of-use or proposal document to extract its principal topics and operational claims. Based on this semantic representation, the module performs relevance-based matching between proposal content and the structured regulatory requirements produced by the RPEM. This coordination ensures that only contextually applicable regulations are forwarded to the compliance reasoning stage.

For each matched regulation–proposal pair, ARCCS employs a retrieval-augmented compliance mechanism that incorporates the most relevant proposal segments into a context-aware decision process. Each regulatory requirement is assigned one of four compliance labels: *Compliant*, *Non-Compliant*, *Insufficient Information*, or *Human Required*. These labels collectively capture clear cases of regulatory satis-

[1]All source code and datasets used in this work are publicly available at: https://github.com/geofila/ARCCS.

faction or violation, as well as situations characterized by missing, ambiguous, or context-dependent information that prevent a fully automated and reliable compliance determination, thereby supporting both automated assessment and human expert review when necessary.

Overall, ARCCS operationalizes regulatory compliance checking as a unified and modular framework that integrates structured legal representation, semantic alignment, and human-in-the-loop decision support within a transparent and explainable technical architecture.

2 Related Work

Recent advances in LLMs have enabled a new class of approaches for supporting regulatory and legal compliance tasks. Prior work has demonstrated that transformer-based architectures can be applied to classify regulatory provisions and assist in automated compliance assessment across domains such as data protection and technical standards, yielding substantial reductions in manual review effort. Complementary systems integrate multiple foundation models into domain-specific software environments to semi-automate the verification of technical artifacts, such as architectural or engineering designs, against formal regulatory constraints, reporting improvements in both throughput and violation detection accuracy.

A growing line of research adopts hybrid and retrieval-augmented paradigms that combine symbolic or structured representations of regulations with LLM-based semantic retrieval and reasoning. In these approaches, regulatory texts are transformed into intermediate abstractions, such as rule sets, graphs, or executable representations, which are then aligned with candidate documents or operational descriptions. This design improves grounding and interpretability while preserving the flexibility of neural language models. Related work has further explored the translation of regulatory clauses into programmatic or semi-formal specifications to enable automated or tool-assisted compliance checking, particularly in financial and technical regulatory settings.

In parallel, substantial progress has been made in the automated analysis of privacy policies. LLM-based frameworks have been shown to achieve strong performance in categorizing policy segments into data practice and user rights taxonomies, as well as in aligning policy text with legal requirements derived from regulations such as the GDPR and U.S. state-level privacy laws. These systems consistently outperform earlier rule-based and keyword-driven baselines, demonstrating the effectiveness of semantic modeling for clause detection and regulatory mapping. A related research direction addresses policy completeness, focusing on whether mandatory regulatory obligations are sufficiently specified in a given document. Empirical studies indicate that a large fraction of real-world policies omit critical legal requirements, motivating the development of semantic and model-driven methods for detecting underspecified or missing obligations.

2.1 Datasets and Benchmarks

The development of compliance-oriented models has been supported by the release of regulation-aware datasets and multi-task benchmarks targeting legal and privacy language understanding. Large-scale corpora of expert-annotated privacy policies have been introduced, where policy segments are labeled according to their correspondence with specific regulatory disclosure requirements. These datasets enable fine-grained evaluation of regulatory alignment and support scalable compliance auditing.

More comprehensive benchmark suites unify multiple legal and privacy tasks, including document classification, sentence- and token-level information extraction, and question answering over regulatory and policy texts. Results on these benchmarks consistently show that domain-adapted models outperform general-purpose language models, highlighting the linguistic and semantic specificity of regulatory discourse. In the broader legal NLP domain, standardized benchmarks cover tasks such as judicial decision classification and contract fairness assessment, while complementary datasets focus on technical regulations, annotating regulatory sentences with entities and relational structures to facilitate rule extraction and machine-interpretable compliance representations. Process-oriented datasets derived from real-world, regulation-driven workflows further support the study of compliance from an operational and auditing perspective.

Despite these advances, existing datasets and methods primarily target isolated subtasks, such as clause classification, regulatory retrieval, information extraction, or completeness checking, and are typically grounded in domain-specific regulatory schemas or task-specific supervision. As a result, they do not directly support the evaluation of compliance as an end-to-end process that begins with unstructured regulatory text and concludes with structured, regulation-level compliance judgments over heterogeneous proposal or terms-of-use documents.

In contrast, the framework introduced in this work, ARCCS, is designed as a general-purpose, modular, and end-to-end LLM-based system that operationalizes regulatory compliance as a unified pipeline. ARCCS integrates hierarchical regulatory extraction, semantic alignment, and retrieval-augmented compliance classification to produce traceable, regulation-level compliance decisions with explicit confidence estimates and structured justifications. This design addresses a gap in the current literature by enabling domain-agnostic compliance assessment across arbitrary regulatory texts and corresponding policy or access documents within a single, coherent evaluation framework.

3 System Overview

This section provides a high-level overview of the ARCCS architecture. As illustrated in Figure 1, the system takes as input a regulatory document and a terms-of-use or proposal document, and produces a regulation-level compliance report with structured justifications and uncertainty-aware labels. The following sections describe the internal design and implementation of the Regulatory Processing and Extraction Module (RPEM) and the Compliance Classification Module (CCM).

3.1 Regulatory Processing and Extraction Module (RPEM)

The Regulatory Processing and Extraction Module (RPEM) is responsible for transforming raw regulatory text into a structured, refined, and machine-interpretable representation of regulatory requirements. This module constitutes the first stage of the ARCCS pipeline and serves as the primary source of regulatory constraints for downstream compliance assessment.

Formally, given a regulatory text D segmented into chunks $\mathcal{C} = \{c_1, \ldots, c_m\}$, the RPEM computes an extraction mapping

$$\phi : \mathcal{C} \to \mathcal{R},$$

where $\mathcal{R} = \{r_1, \ldots, r_n\}$ is the set of structured regulatory requirement objects.

The RPEM operates on regulatory documents that have been preprocessed into a normalized textual format. The module performs hierarchical segmentation of the regulation by decomposing the text into legally meaningful units, including articles, paragraphs, and subparagraphs. Each resulting regulatory chunk is treated as an atomic unit of analysis and is associated with structural metadata, such as its article identifier and positional context within the regulation. This design preserves local legal context while enabling fine-grained downstream processing.

For each regulatory chunk, the RPEM applies a structured extraction procedure to identify candidate regulatory requirements. This process determines the legal function of the provision (e.g., obligation, right, prohibition, or condition), the applicable legal actor, and any associated constraints or qualifiers. The extracted information is stored in a structured representation that links each requirement to its source text and structural metadata, thereby supporting traceability and explainability in subsequent compliance decisions.

Regulation Filtering and Deduplication Not all regulatory chunks yield actionable compliance constraints. The RPEM therefore applies a filtering stage to remove provisions that are primarily definitional, descriptive, or contextual in nature and do not impose enforceable obligations or rights. This step reduces noise in the regulatory representation and ensures that downstream components operate on a set of operationally relevant requirements.

Regulatory texts frequently contain overlapping or semantically equivalent provisions expressed across multiple sections or through cross-references. To address this, the RPEM performs a deduplication and consolidation procedure that identifies structurally or semantically similar regulatory requirements and

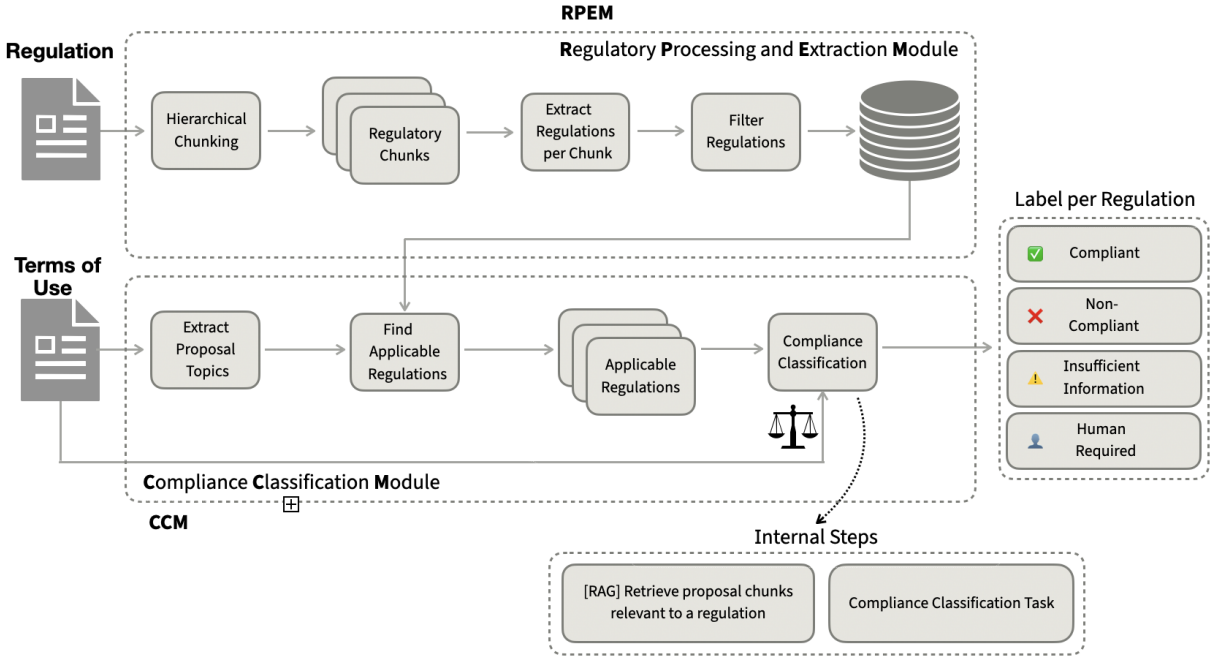

Figure 1: Overview of the proposed regulatory compliance assessment pipeline.

merges them into unified representations. This process reduces redundancy and produces a compact set of distinct regulatory constraints while preserving references to all source provisions.

The output of the RPEM is a refined collection of structured regulatory requirements, each represented by a unique identifier, a formalized description of the regulatory constraint, associated legal metadata, and a set of references to the original regulatory text. Specifically, encodes the law document in the form of a list of structured, article-level regulatory objects, where each element represents a single, atomic, and operationally meaningful legal requirement. Each object encodes the legal function of the provision, the regulated actor, the jurisdictional and domain scope, the set of mandatory obligations, and explicit references to the authoritative source text. This representation serves as a normalized regulatory knowledge layer that enables fine-grained applicability analysis and regulation-level compliance reasoning in the CCM.

Formally, the RPEM output is defined as:

$$\mathcal{R} = \{r_1, r_2, \ldots, r_n\},$$

where each r_i is a structured regulatory object consisting of semantic, legal, and traceability fields.

3.2 Compliance Classification Module (CCM)

The Compliance Classification Module (CCM) is responsible for determining whether the structured regulatory requirements produced by the RPEM are applicable to, and subsequently satisfied by, the proposal document. This module constitutes the second stage of the ARCCS pipeline and performs regulation-level compliance assessment through applicability analysis, semantic alignment, retrieval-augmented reasoning, and uncertainty-aware labeling.

For each regulatory requirement, the CCM first evaluates whether the requirement is applicable to the proposal under analysis. Applicability is determined based on a set of contextual parameters, including the regulatory scope and domain, the type of system or service described in the proposal, and jurisdictional or geographical constraints (e.g., whether the regulation explicitly applies to the country or region in which the system is deployed). Regulatory requirements that are deemed non-applicable are excluded from further compliance reasoning, thereby reducing analytical noise and improving decision relevance.

For applicable regulatory requirements, the CCM processes the proposal or terms-of-use document to extract its principal topics and

operational claims, forming a semantic representation of the system's stated practices, commitments, and behaviors. Based on this representation, the module performs relevance-based matching to identify the subset of proposal segments that are semantically related to each regulatory requirement. This alignment step ensures that compliance reasoning is grounded in contextually relevant evidence and preserves traceability between legal constraints and proposal content.

Let the proposal document be segmented into passages $\mathcal{P} = \{p_1, \ldots, p_k\}$. For each requirement $r_i \in \mathcal{R}$, the CCM computes a relevance score function

$$\rho(r_i, p_j) \in [0, 1],$$

and selects evidence

$$E_i = \{p_j \in \mathcal{P} : \rho(r_i, p_j) \geq \tau\},$$

where τ is a matching threshold.

For each matched regulation–proposal pair, the CCM employs a retrieval-augmented mechanism that incorporates the most relevant proposal segments into a context-aware compliance reasoning process. This design supports explainable and evidence-based assessment by conditioning each compliance decision on explicitly retrieved textual support, thereby reducing the likelihood of unsupported or hallucinated judgments.

For each regulatory requirement, the CCM assigns both a compliance label and a corresponding confidence score that reflects the system's estimated reliability of the decision. The confidence score is derived from factors such as the consistency of the retrieved evidence, the completeness of the proposal information with respect to the regulatory conditions, and the clarity of the semantic alignment.

Each regulatory requirement is assigned one of the following compliance labels:

- **Compliant**: Indicates that the proposal explicitly satisfies the regulatory requirement. All mandatory conditions specified by the regulation are addressed in the retrieved proposal segments, and no conflicting statements are identified. The assigned confidence score reflects strong and consistent supporting evidence.
- **Non-Compliant**: Indicates a clear and explicit violation of the regulatory requirement. The proposal contradicts a mandatory legal constraint or omits a required condition in a manner that constitutes a definitive regulatory mismatch. This label is associated with high-confidence evidence of non-conformance.
- **Insufficient Information**: Indicates that the proposal references the relevant regulatory topic but does not provide sufficient detail to determine compliance. Critical conditions, qualifiers, or dependencies required by the regulation are missing from the retrieved proposal segments, resulting in a low or moderate confidence score.
- **Human Required**: Indicates that an automated compliance determination cannot be reliably made. This label is assigned in cases of ambiguity, contradictory statements, missing external legal or technical context, or when the model confidence for the predicted label falls below a predefined threshold. In such cases, the assessment is explicitly deferred to a human expert for final evaluation.

Compliance Report Generation. The output of the CCM is a regulation-level compliance report in which each regulatory requirement is associated with an applicability decision, a compliance label, a confidence score, and a structured explanation. The explanation explicitly states the rationale for the assigned label, including which regulatory conditions were satisfied or violated, or which information was missing or ambiguous in the proposal. This design supports transparency, auditability, and effective human-in-the-loop review in real-world regulatory compliance workflows.

4 Experimental Setup

This section reports two complementary evaluation settings. First, ARCCS is evaluated on a realistic compliance-checking setting in which a terms-of-use document is assessed against requirements extracted from the EU General Data Protection Regulation (GDPR)[2]. Second, we introduce a quantitative event-log

[2] https://gdpr-info.eu

compliance benchmark for public procurement grounded in Directive 2014/24/EU (CELEX: 32014L0024). Across both settings, we focus on interpretable, regulation-level outcomes and traceable evidence.

4.1 Policy Document Compliance Setup

Inputs. We used the GDPR as the regulatory input and evaluated three widely used terms-of-use documents: *WhatsApp Terms of Service*[3], *Netflix Terms of Use*[4], and *ChatGPT Terms of Use*[5].

Regulatory requirements. Given the raw GDPR text, the RPEM segmented the regulation and extracted candidate requirements. After filtering non-actionable clauses and applying deduplication, the system produced $|\mathcal{R}| = 100$ atomic regulatory requirements.

Compliance assessment. For each $r_i \in \mathcal{R}$, the CCM matched the most relevant document evidence and produced a regulation-level label (*Compliant*, *Non-Compliant*, *Insufficient Information*, or *Human Required*) with an accompanying explanation. The confidence threshold for triggering the *Human Required* label was set to 70%.

Model configuration. Unless otherwise stated, experiments were run using GPT-5.2[6].

Evaluation protocol. We report the distribution of predicted compliance labels across the 100 extracted requirements for each document. In addition, we employed an *LLM-as-a-judge* protocol (Gu et al., 2025) to evaluate system outputs at the regulation level: for each requirement, we provided the judge model with (i) the extracted regulation text/requirement representation, (ii) the relevant terms-of-use document context, and (iii) the full ARCCS output for that requirement (predicted label and explanation). The judge then assessed whether the proposed compliance decision and its justification are consistent with the provided regulation and evidence (see Appendix B). For the procurement benchmark, evaluation is fully quantitative with deterministic ground-truth labels at both rule and case level (described below in this setup).

Sampled evaluation set. To obtain a quantitative estimate of decision quality, we additionally performed a small-scale, sampled evaluation: we drew a sample of 30 regulation-level checks from each of the three terms-of-use documents (WhatsApp, Netflix, and ChatGPT) under the same GDPR regulatory context, yielding 90 judged instances in total.

Evaluator models. We used three model variants as evaluators: gpt-5.1[7], gpt-5.2[8], and gpt-5.2-pro[9]. For gpt-5.2, the reasoning effort was set to *medium*. Quantitative accuracy and inter-evaluator agreement results are reported in Section 5.

4.2 Public Procurement Quantitative Setup

Legal scope and target rules. In addition to the GDPR-based terms-of-use setting, we evaluate an LLM-based log compliance checking setup for public procurement grounded in Directive 2014/24/EU on public procurement (CELEX: 32014L0024). The Directive defines the core EU legal framework for public contract award procedures, including value thresholds, publication and transparency obligations, principles for participant selection and award, and lifecycle conditions for contract execution and termination. Rather than treating the full Directive as a monolithic target, we evaluate a focused subset of twelve rules derived from Articles 4, 48, 56, and 73 as represented in the extracted regulatory knowledge base. The selected rules capture four compliance dimensions: monetary constraints, duplicate-publication restrictions, temporal/logical consistency of award procedures, and lifecycle consistency for contract execution and termination.

Concretely, the twelve rules cover:

1. maximum contract amount threshold,
2. prohibition of duplicate publication of the same call,

[3] https://www.whatsapp.com/legal/terms-of-service/revisions/20210104

[4] https://help.netflix.com/legal/termsofuse?locale=en-GB

[5] https://openai.com/en-GB/policies/row-terms-of-use

[6] gpt-5.2-2025-12-11

[7] gpt-5.1-2025-11-13

[8] gpt-5.2-2025-12-11

[9] gpt-5.2-pro-2025-12-11

3. prohibition of award before publication,
4. prohibition of award before participation,
5. requirement that publication and participation eventually lead to an award,
6. maximum delay of 70 days between publication and award,
7. prohibition of contract start without prior award,
8. prohibition of contract end before publication,
9. prohibition of contract end immediately after publication without participation and award,
10. requirement that any started contract must eventually have an end date,
11. requirement that any ended contract must previously have started,
12. requirement that any terminated contract must previously have been awarded.

Event-log data source. The empirical context is based on procurement event logs derived from TED (Tenders Electronic Daily), the official EU platform for public procurement notices. TED provides large-scale, structured, real-world procurement traces and is therefore suitable for process-centric compliance analysis. Each procurement case is represented by a unique `case_id` and a temporally ordered trace including events such as `PUBLICATION`, `PARTICIPATION`, `AWARD`, `CONTRACT-START`, and `CONTRACT-END`. In addition to event type and timestamp, each log entry includes procurement-relevant attributes (e.g., contract amount, procedure type, electronic handling indicator, framework-agreement indicator, country, NUTS code, CPV division, CPV code, and trace length). This structure is essential because the target legal constraints depend on event order, event co-occurrence, missing lifecycle steps, and attribute values.

Controlled benchmark construction. To enable balanced and reproducible evaluation across all twelve rules, we construct a controlled benchmark of 100 procurement traces using the same schema as TED logs. This is necessary because several violations of interest are sparse in naturally occurring subsets of procurement data. The benchmark includes both compliant and non-compliant traces, with positive and negative examples deliberately distributed across rules. Each synthetic case preserves the event vocabulary and metadata fields of TED-style data, while enabling explicit ground-truth assignment.

For every case, expected labels are derived independently through deterministic logic over event sequences and attributes. For example, if an `AWARD` event precedes `PUBLICATION`, the corresponding rule is violated; if a trace contains `CONTRACT-START` but no `CONTRACT-END`, the lifecycle-completion rule is violated; if contract amount exceeds the legal threshold, the monetary rule is violated. This ensures that model outputs are evaluated against externally defined legal-process logic rather than self-referential model judgments.

Evaluation granularity and metrics. The benchmark contains 100 distinct procurement traces, each identified by a unique `case_id`, and each trace is evaluated against all twelve targeted regulations. This yields 1,200 rule-level decisions in total (100 traces $\times$ 12 rules), since every trace produces one compliance prediction per rule. To ensure meaningful class coverage, the dataset includes both violating and non-violating examples: 60 traces are constructed to trigger at least one targeted violation, while 40 traces serve as fully compliant controls. Moreover, even within violating traces, only a subset of the twelve rules is positive and the remaining rules are negative, ensuring that rule-level evaluation includes both contradiction (violation) and non-contradiction (non-violation) outcomes.

Evaluation is performed at two levels. At rule level, the system produces one compliance decision per `(case, rule)` pair. At case level, rule outputs are aggregated into a binary decision indicating whether a procurement trace contains at least one violation. For each case, the system stores the raw trace, expected labels, per-rule outputs, and the final case-level decision, enabling full post hoc error analysis.

Performance is quantified with standard binary classification metrics, where the positive

class corresponds to true legal violation detection, operationalized as a `NON_COMPLIANT` prediction. We report accuracy, precision, recall, and F1-score at rule level, plus case-level metrics obtained by collapsing the twelve rule outputs into a single violation/no-violation label per trace. We additionally report exact match, defined as the proportion of cases for which all twelve rule decisions are simultaneously correct.

Methodological objective. This quantitative setup combines a formal legal basis (Directive 2014/24/EU), realistic TED-style process traces, deterministic external ground truth, and multi-level metrics. The resulting protocol enables rigorous analysis of LLM-based compliance checking for procurement event logs in terms of overall performance, per-rule behavior, and full-case exact consistency.

Model variants. For the procurement benchmark, we evaluated three model variants under the same legal scope and evaluation protocol: GPT-5.4, GPT-5-mini, and GPT-5.2. The rest of paremeters was the same as the previous one.

Unless explicitly stated otherwise, the remaining experimental parameters were kept the same as in the previous setup.

5 Results

We report results for two complementary experiments: (i) GDPR terms-of-use compliance assessment with LLM-as-a-judge validation and qualitative contradiction analysis, and (ii) quantitative public procurement log compliance checking under Directive 2014/24/EU.

5.1 Results on Policy Document Compliance

Table 1 reports the accuracy of three evaluator variants on the sampled GDPR set. High accuracy indicates that ARCCS decisions and explanations are generally judged to be consistent with the provided legal text and retrieved evidence. More importantly, these scores suggest that the system produces evidence-grounded rationales rather than unsupported label assignments.

Table 2 reports inter-evaluator agreement. We observe substantial overall agreement (Fleiss' $\kappa = 0.69$, $p < 0.001$). In particular, GPT-5.2 and GPT-5.2-pro exhibit perfect agreement ($\kappa = 1.00$, $p < 0.001$), indicating stable evaluator behavior under this setup.

Evaluator (LLM-as-a-Judge)	Accuracy (%)
GPT-5.1	96.67
GPT-5.2	90.00
GPT-5.2-pro	90.00

Table 1: Accuracy of model-based evaluators on the sampled set (30 checks per terms-of-use document; 90 total).

Agreement	κ	p-value
Cohen's κ (GPT-5.1 vs GPT-5.2)	0.47	0.19
Cohen's κ (GPT-5.1 vs GPT-5.2-pro)	0.47	0.19
Cohen's κ (GPT-5.2 vs GPT-5.2-pro)	1.00	< 0.001
Fleiss' κ (3 raters)	0.69	< 0.001

Table 2: Inter-evaluator agreement on the binary verdict (YES/NO) for the sampled set.

Taken together, the accuracy and agreement statistics indicate that ARCCS outputs are internally coherent and evidence-aligned: evaluator models typically concur that the predicted label follows from both regulation text and retrieved supporting evidence.

Label Distribution and Contradiction Analysis. The compliance label distributions produced by ARCCS are summarized in Table 3. Across documents, *Insufficient Information* is the dominant outcome, consistent with the fact that many GDPR obligations depend on internal governance artifacts that are usually absent from public terms-of-use pages.

To contextualize label differences, Table 4 reports document size statistics. The substantially longer WhatsApp terms likely improve evidence coverage and yield more determinate outcomes than shorter documents.

Beyond underspecification, ARCCS also surfaces explicit legal contradictions. Below we present an illustrative contradiction extracted from the *ChatGPT Terms of Use*.

Contradiction Case Study (GDPR Article 79)

Compliance Status: NON-COMPLIANT
Regulation: General Data Protection Regulation (EU) 2016/679, Article 79: Right to an effective judicial remedy against a controller or processor
Regulation ID: GDPR Article 79
Domain: General

Document	Compliant	Non-Compliant	Insufficient Info.	Human Req.
WhatsApp	2	2	89	7
Netflix	0	0	99	1
ChatGPT	0	4	94	2

Table 3: Compliance label distribution over 100 GDPR-derived regulatory requirements for each evaluated terms-of-use document.

Document	Characters	Words	Sentences	Paragraphs	Lines
WhatsApp	323682	51698	1862	1382	3723
Netflix	15777	2538	128	39	158
ChatGPT	20404	3324	171	100	197

Table 4: Terms-of-use document sizes used in our experiments.

Contradiction Details. Document mandates arbitration and exclusive California courts for claims, which can deny/limit EU/EEA data subjects' ability to bring GDPR-rights infringement proceedings before courts in their habitual residence or the controller's EU establishment as required by GDPR Article 79.
Conflicting text includes: (1) mandatory arbitration requirement for disputes, and (2) exclusive forum selection in San Francisco, California for claims (except as provided in arbitration section). These provisions contradict the GDPR Article 79 requirement that data subjects must have access to an effective judicial remedy and be able to bring proceedings in the Member State of their habitual residence (subject to the public-authority exception).

Evidence from Document.

> "MANDATORY ARBITRATION. You and OpenAI agree to resolve any claims arising out of or relating to these Terms or our Services... through final and binding arbitration."
>
> "Governing law. California law will govern these Terms... Except as provided in the dispute resolution section above, all claims arising out of or relating to these Terms will be brought exclusively in the federal or state courts of San Francisco, California."

Explanation. GDPR Article 79 provides data subjects a right to an effective judicial remedy against a controller/processor and allows proceedings to be brought in the Member State of the controller/processor's establishment or the data subject's habitual residence (with a limited exception for public authorities acting in public powers). The Terms impose (i) final and binding arbitration for "any claims arising out of or relating to these Terms or our Services" and (ii) an exclusive court forum in San Francisco, California for claims outside arbitration. These clauses, as written, restrict access to EU Member State courts and can functionally deny the Article 79 venue rights and judicial remedy in EU courts for GDPR infringements. The document does not carve out GDPR/data-protection claims for EU/EEA users to preserve Article 79 rights; instead it broadly applies arbitration and a non-EU exclusive forum, creating a direct conflict with the regulation's required availability of EU judicial remedies and jurisdiction options.

This case study illustrates ARCCS's ability to move beyond missing-information detection and identify explicit normative conflicts between contractual provisions and legal obligations.

5.2 Results on Public Procurement Logs

This experiment evaluates regulation-level compliance prediction on 100 synthetic public-procurement traces derived from Directive 2014/24/EU. Each trace is checked against 12 legal constraints, yielding 1,200 rule-level decisions per model. The benchmark is intentionally demanding because multiple constraints depend on event ordering and lifecycle completeness; therefore, maintaining consistency across all rule decisions is harder than simply identifying whether a trace is problematic at high level.

Tables 5, 6, and 7 summarize rule-level performance, document-level screening performance, and strict all-rule correctness across GPT-5.4[10], GPT-5-mini[11], and GPT-5.2.

5.2.1 Rule-Level Performance

At rule level, all three models perform strongly and remain tightly clustered. Accuracy ranges from 98.5% to 98.8%, while recall remains near-saturated (98.8%–100.0%), indicating robust detection of true violations. GPT-5-mini and GPT-5.2 jointly provide the strongest ag-

[10] gpt-5.4-2026-03-05
[11] gpt-5-mini-2025-08-07

Model	Acc.	Prec.	Rec.	F1
GPT-5.4	98.5	83.2	98.8	90.3
GPT-5.2	98.8	85.9	100.0	92.4
GPT-5-mini	98.8	85.9	100.0	92.4

Table 5: Rule-level model comparison on the procurement compliance benchmark (100 traces, 1,200 rule decisions per model).

Model	Acc.	Prec.	Rec.	F1
GPT-5.4	98.0	96.8	100.0	98.4
GPT-5.2	100.0	100.0	100.0	100.0
GPT-5-mini	100.0	100.0	100.0	100.0

Table 6: Document-level model comparison on the procurement benchmark.

gregate rule-level profile, matching on precision (85.9%), recall (100.0%), and F1 (92.4%), and outperforming GPT-5.4 primarily through fewer false alarms. Overall, the residual error profile remains false-positive-driven rather than miss-driven, which is consistent with the near-perfect recall across models.

5.2.2 Document-Level Performance

At document level (violation/no-violation per trace), performance is extremely strong. GPT-5.4 reaches 98.0% accuracy and 98.4% F1, correctly identifying all violating traces and 95% of compliant traces. GPT-5-mini and GPT-5.2 achieve 100.0% on all document-level metrics. These results show that ARCCS is highly effective as a first-stage screening system for procurement workflows: in practical terms, it reliably separates problematic from compliant traces. This is particularly important in compliance screening, where reliably detecting problematic cases is often more critical than eliminating every residual false alarm.

5.2.3 Exact-Match Performance

To evaluate strict trace-level correctness, we report exact match, where a trace is counted as correct only if all 12 rule-level decisions are simultaneously correct. Under this criterion, scores are lower than document-level metrics because a single rule-level mismatch is sufficient for a non-exact trace. This behavior is expected in multi-rule compliance settings and should be interpreted as a fine-grained calibration challenge rather than a failure of core violation detection.

Overall, the comparative pattern is clear: all three models are highly reliable for regulation-level retrieval and near-perfect for document-level screening, while the remaining headroom lies in achieving strict all-rules-correct predictions. GPT-5-mini and GPT-5.2 are effectively tied on aggregate rule-level and document-level metrics, but GPT-5-mini is marginally better on exact match (87.0% vs 86.0%), indicating slightly better fine-grained calibration when all 12 rules must be simultaneously correct. The gap between near-perfect document-level scores and lower exact-match scores is primarily explained by isolated extra false alarms on individual rules rather than broad misunderstanding of entire traces. From a practical perspective, GPT-5-mini provides the strongest balance in this benchmark, combining top aggregate performance with the best exact-match rate and a lower *Insufficient Information* rate (2.3%) compared with GPT-5.2 (7.3%) and GPT-5.4 (8.8%). At the same time, the uniformly high scores across all three models reinforce that a substantial share of performance comes from the ARCCS architecture itself, which remains robust across model backbones.

Model	Exact Match (%)
GPT-5.4	83.0
GPT-5.2	86.0
GPT-5-mini	87.0

Table 7: Document-level exact-match performance, i.e., the percentage of traces for which all 12 rule decisions are simultaneously correct.

6 Demo Interface

To support adoption by non-technical stakeholders, we implemented a lightweight web-based demo for ARCCS (Figure 2) in the policy document compliance use case. The interface is designed to minimize user effort and to abstract away model configuration and pipeline orchestration.

The demo follows a simple three-step workflow: (i) upload a regulatory document (or select a preloaded regulation such as the GDPR), (ii) upload a target policy/terms-of-use document, and (iii) run the compliance analysis. During execution, the demo streams real-time logs to provide transparency into pipeline progress and intermediate stages. After completion, the system generates a structured compliance report and maintains a history of prior runs to support iterative auditing and compar-

ison across documents.

7 Experimental Resources Used

The experimental workflow was implemented using a development and prototyping environment that integrated a broad range of open-source and proprietary foundation models, including open-weight families such as Qwen[12], Mistral[13], and EuroLLM[14], as well as models tailored for the legal domain such as Saul (Colombo et al., 2024)[15]. In addition, Amazon Bedrock and OpenAI were employed to provide access to a broad range of closed commercial models and to offer unified access to multiple model families for large-scale comparative experimentation. This configuration enabled systematic evaluation across diverse model families and established a widely adopted, standardized reference point for assessing reasoning quality, stability, and output consistency under identical task and prompt configurations.

This phase focused on validating the end-to-end pipeline, refining prompt design, and assessing the stability and suitability of different model architectures for long-context regulatory and legal document processing. In particular, the availability of diverse model families through Amazon Bedrock supported broad comparative experimentation across architectures and parameter scales, while OpenAI models were used alongside this process as widely adopted reference models, providing a common and standardized benchmark for assessing reasoning quality and output consistency under identical task and prompt configurations.

For the final experimental evaluation, we employed OpenAI models, which exhibited more stable reasoning patterns and higher reliability when operating on complex, multi-step compliance reasoning tasks and extended legal inputs. This configuration further enabled the use of a consistent model family for both system inference and LLM-based evaluation, thereby reducing cross-model bias and improving the reproducibility of regulation-level compliance assessments. From a resource allocation perspective, this staged workflow supported broad, cost-efficient model exploration during development, followed by targeted, high-fidelity experimentation in the final evaluation phase, ensuring that computational resources were concentrated on configurations that demonstrated the strongest empirical performance and methodological robustness.

[12] https://huggingface.co/collections/Qwen/qwen3

[13] https://huggingface.co/docs/transformers/en/model_doc/mistral

[14] https://huggingface.co/blog/eurollm-team/eurollm-9b

[15] https://huggingface.co/papers/2403.03883

8 Conclusion

This work introduced ARCCS, a modular and transparent framework for automated, regulation-level compliance assessment that integrates structured regulatory extraction, semantic alignment, and retrieval-augmented reasoning within a unified multi-agent architecture. By transforming unstructured legal texts and policy documents into traceable, machine-interpretable representations, ARCCS enables scalable compliance evaluation while preserving explainability and human oversight through uncertainty-aware labeling and structured justifications.

Experimental results on GDPR-based compliance checking demonstrate that the system produces coherent, evidence-aligned decisions across diverse terms-of-use documents, with substantial agreement among independent LLM-based evaluators. Beyond underspecification detection, ARCCS is capable of surfacing concrete regulatory contradictions, highlighting its practical value for auditing and legal review workflows. In parallel, we define a dedicated quantitative benchmark for procurement event-log compliance under Directive 2014/24/EU, enabling deterministic rule-level and case-level assessment in process-centric public contracting scenarios. Empirically, this benchmark shows high and stable performance across model variants, with rule-level accuracy in the 98.4%–98.8% range and document-level F1 in the 98.4%–100.0% range, indicating robust violation screening under process constraints. Exact-match results (83.0%–87.0%) further show that the main remaining challenge is full cross-rule consistency rather than core detection quality, reinforcing the robustness of the ARCCS architecture.

As a future direction, we plan to conduct a

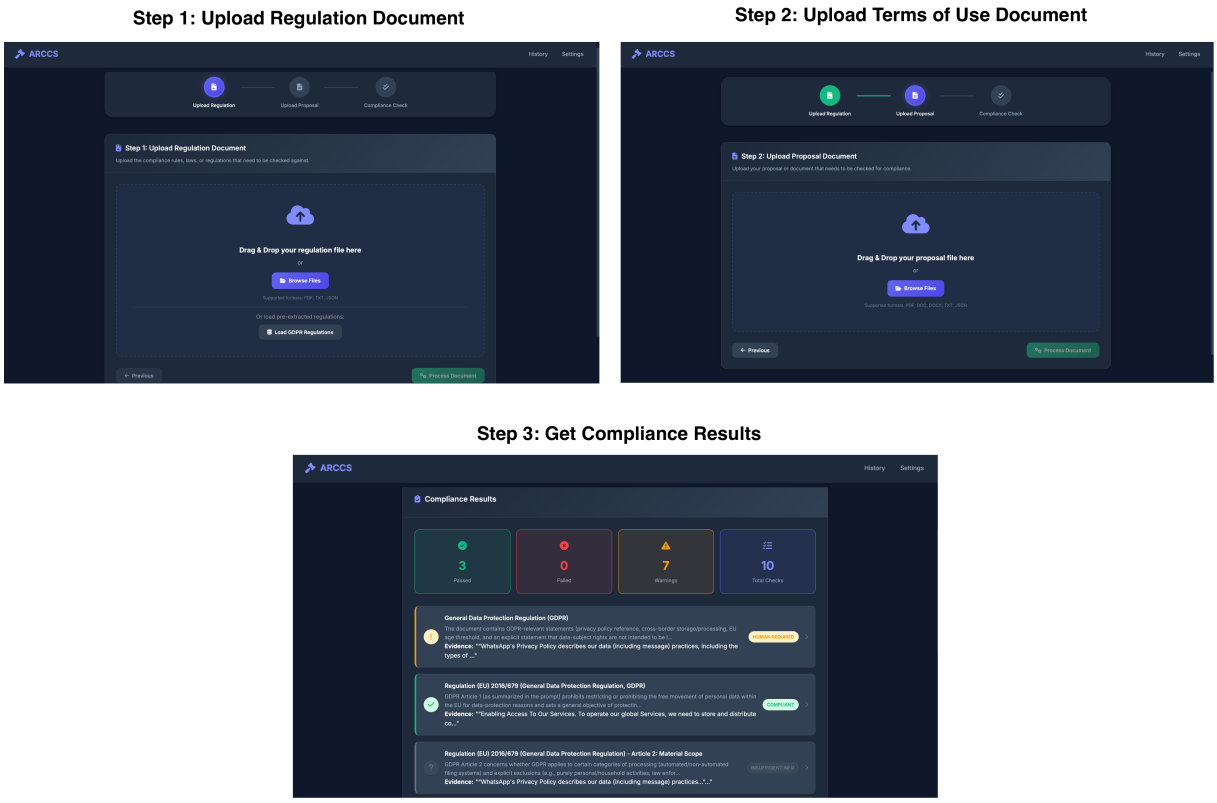

Figure 2: Web demo interface for ARCCS, designed for users without programming experience.

more rigorous human-centered evaluation involving legal and domain experts to assess decision quality, interpretability, and practical usability in real-world compliance workflows. In parallel, we aim to adapt and deploy ARCCS within specific application domains and public-sector institutions, enabling context-aware customization of regulatory representations and compliance criteria to meet the operational and jurisdictional requirements of governmental and institutional settings.

Acknowledgments

This work was fully funded by the 'OptiGov' project, with ref. n. 2024.07385.IACDC (DOI: 10.54499/2024.07385.IACDC), fully funded by the 'Plano de Recuperação e Resiliência' (PRR) under the investment 'RE-C05-i08 - Ciência Mais Digital' (measure 'RE-C05-i08.m04'), framed within the financing agreement signed between the 'Estrutura de Missão Recuperar Portugal' (EMRP) and Fundação para a Ciência e a Tecnologia, I.P. (FCT) as an intermediary beneficiary.

References

Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4402–4417, Online. Association for Computational Linguistics.

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749, Online. Association for Computational Linguistics.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In

Proceedings of the 24th International Conference on Machine Learning, pages 33–40.

Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110.

Orlando Amaral Cejas, Sallam Abualhaija, and Lionel C. Briand. 2024. Compai: A tool for gdpr completeness checking of privacy policies using artificial intelligence. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ASE '24, page 2366–2369, New York, NY, USA. Association for Computing Machinery.

Orlando Amaral Cejas, Sallam Abualhaija, Nicolas Sannier, Marcello Ceci, and Domenico Bianculli. 2025. Gdpr compliance in privacy policies of mobile apps: An overview of the state-of-practice. In *2025 IEEE 33rd International Requirements Engineering Conference (RE)*, pages 320–331.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Hao Chen, Quan Liu, Ke Fu, Jian Huang, Chang Wang, and Jianxing Gong. 2022. Accurate policy detection and efficient knowledge reuse against multi-strategic opponents. *Knowledge-Based Systems*, 242:108404.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

VV Denisov, Elena Belkina, and Dirk Fahland. 2018. Bpic'2018: Mining concept drift in performance spectra of processes. In *8th International Business Process Intelligence Challenge*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Shabnam Hassani. 2024. Enhancing Legal Compliance and Regulation Analysis with Large Language Models . In *2024 IEEE 32nd International Requirements Engineering Conference (RE)*, pages 507–511, Los Alamitos, CA, USA. IEEE Computer Society.

Hansi Hettiarachchi, Amna Dridi, Mohamed Medhat Gaber, Pouyan Parsafard, Nicoleta Bocaneala, Katja Breitenfelder, Gonçal Costa, Maria Hedblom, Mihaela Juganaru-Mathieu, Thamer Mecharnia, and 1 others. 2025. Codeaccord: A corpus of building regulatory data for rule generation towards automatic compliance checking. *Scientific data*, 12(1):170.

Siyuan Li, Jian Chen, Rui Yao, Xuming Hu, Peilin Zhou, Weihua Qiu, Simin Zhang, Chucheng Dong, Zhiyao Li, Qipeng Xie, and Zixuan Yuan. 2026. Compliance-to-code: Enhancing financial compliance checking via code generation. *Preprint*, arXiv:2505.19804.

Soumya Madireddy, Lu Gao, Zia Din, Kinam Kim, Ahmed Senouci, Zhe Han, and Yunpeng Zhang. 2025. Large language model-driven code compliance checking in building information modeling. *Preprint*, arXiv:2506.20551.

Maaz Bin Musa, Steven M. Winston, Garrison Allen, Jacob Schiller, Kevin Moore, Sean Quick, Johnathan Melvin, Padmini Srinivasan, Mihailis E. Diamantis, and Rishab Nithyanand. 2024. C3PA: An open dataset of expert-annotated and regulation-aware privacy policies to enable scalable regulatory compliance audits. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3710–3722, Miami, Florida, USA. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Atreya Shankar, Andreas Waldis, Christof Bless, Maria Andueza Rodriguez, and Luca Mazzola. 2023. Privacyglue: A benchmark dataset for general language understanding in privacy policies. *Applied Sciences*, 13(6).

Jingyun Sun, Zhongze Luo, and Yang Li. 2025. A compliance checking framework based on retrieval augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2603–2615, Abu Dhabi, UAE. Association for Computational Linguistics.

Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu, Dajiang Zhu, Quanzheng Li, Xiang Li, Tianming Liu, and Lei Fan. 2023.

Policygpt: Automated analysis of privacy policies with large language models. *Preprint*, arXiv:2309.10238.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Qinge Xie, Karthik Ramakrishnan, and Frank Li. 2025. Evaluating privacy policies under modern privacy laws at scale: an llm-based automated approach. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA. USENIX Association.

A Example Module Outputs (GDPR Case Study)

This appendix provides a concrete example of the structured outputs produced by the RPEM and CCM modules. The example is based on a regulation from the European Union General Data Protection Regulation (GDPR) [16] and a corresponding compliance assessment performed on a terms-of-use document. The purpose of this appendix is to illustrate the internal data representations exchanged between modules and to demonstrate how traceability and explainability are preserved throughout the pipeline.

A.1 Example Output of the Regulatory Processing and Extraction Module (RPEM)

The RPEM outputs a list of article-level regulatory objects, each representing a single operational legal requirement derived from a regulatory source. An example object corresponding to GDPR Article 5 is shown below.

RPEM Output: Article-Level Regulatory Representation (GDPR)

```
{
  "regulation_id": "GDPR Article 5",
  "regulation_name": "General Data
  ↪ Protection Regulation (EU)
  ↪ 2016/679  Principles relating to
  ↪ processing of personal data",
  "regulation_type": "article",

  "jurisdiction": {
    "geographic_scope": "EU",
    "applicable_regions": ["European
    ↪ Union", "European Economic
    ↪ Area"],
    "cross_border_applicability": true
  },

  "domain": {
    "primary_domain": "Data Protection",
    "sub_domains": [
      "Personal Data Processing",
      "Data Security",
      "Data Minimization",
      "Transparency",
      "Accountability"
    ]
  },

  "legal_function": "obligation",
  "regulated_actor": "data_controller",

  "requirements": {
    "mandatory_obligations": [
      "Process personal data lawfully,
      ↪ fairly, and transparently",
      "Collect data for specified,
      ↪ explicit, and legitimate
      ↪ purposes",
      "Limit personal data collection
      ↪ to what is necessary",
      "Ensure personal data accuracy
      ↪ and allow rectification",
      "Limit storage duration to what
      ↪ is necessary",
      "Ensure integrity and
      ↪ confidentiality through
      ↪ technical and organizational
      ↪ measures",
      "Demonstrate compliance with all
      ↪ processing principles
      ↪ (accountability)"
    ]
  },

  "scope": {
    "applies_when": [
      "Personal data of natural persons
      ↪ is processed",
      "Processing is automated or part
      ↪ of a structured filing
      ↪ system"
    ],
    "exceptions": [
      "Purely personal or household
      ↪ activities",
      "Law enforcement processing under
      ↪ Directive (EU) 2016/680"
    ]
  },

  "source_references": {
    "regulation": "Regulation (EU)
    ↪ 2016/679",
    "article": "Article 5",
```

[16] https://gdpr-info.eu

```
    "official_title": "Principles
    ↪  relating to processing of
    ↪  personal data"
  },

  "traceability": {
    "source_section": "GDPR Article 5",
    "cross_referenced_articles":
    ↪  ["Article 6", "Article 24",
    ↪  "Article 32"]
  },

  "_quality_score": {
    "score": 92.4,
    "status": "KEEP",
    "completeness": "HIGH",
    "issues": []
  }
}
```


A.2 Example Output of the Compliance Classification Module (CCM)

The CCM outputs a list of regulation-level compliance decision objects, each corresponding to a single structured regulatory requirement produced by the RPEM. Each decision object captures the applicability of the regulation to the analyzed document, the assigned compliance label, a confidence estimate, and an evidence-grounded explanation that links regulatory constraints to retrieved segments of the proposal text.

An example compliance decision is shown below for the *WhatsApp Terms of Service (2021)*[17]. In this case, the CCM evaluates the document against multiple provisions of the General Data Protection Regulation (GDPR) and produces both compliant and non-compliant outcomes. The example illustrates the module's ability to identify explicit regulatory violations, such as jurisdictional scope conflicts under GDPR Article 3, while simultaneously confirming compliance with other regulatory principles, including data processing and transparency requirements under GDPR Article 5.

This example demonstrates how the CCM integrates applicability analysis, retrieval-augmented evidence selection, and structured reasoning to generate transparent, regulation-level compliance assessments that are suitable for expert validation and regulatory auditing.

[17] https://www.whatsapp.com/legal/terms-of-service/revisions/20210104

CCM Output: Regulation-Level Compliance Decision


```
{
  "overall_status": "NON-COMPLIANT - 1
  ↪  violation(s) found",
  "summary": {
    "compliant": 19,
    "non_compliant": 1,
    "total": 20,
    "insufficient_info": 0,
    "human_required": 0
  },
  "total_checked": 20,

  "violations": [
    {
      "regulation_id": "GDPR Article 3",
      "regulation_name": "General Data
      ↪  Protection Regulation (EU)
      ↪  2016/679  Territorial scope",
      "contradiction_found": true,
      "compliance_status":
      ↪  "NON_COMPLIANT",
      "contradiction_details": "The
      ↪  document implies an attempt
      ↪  to avoid being subject to
      ↪  foreign jurisdictions by
      ↪  limiting service
      ↪  availability. This conflicts
      ↪  with GDPR Article 3, which
      ↪  establishes that GDPR applies
      ↪  when processing targets
      ↪  individuals in the EU,
      ↪  regardless of the physical
      ↪  location of the service or
      ↪  processing infrastructure.",
      "evidence": "Our Services are not
      ↪  intended for distribution to
      ↪  or use in any country or
      ↪  territory where such use
      ↪  would subject us to
      ↪  regulations in another
      ↪  country or territory."
    }
  ],

  "detailed_results": [
    {
      "regulation_id": "GDPR Article 5",
      "regulation_name": "GDPR
      ↪  Principles relating to
      ↪  processing of personal data",
      "contradiction_found": false,
      "compliance_status": "COMPLIANT",
      "evidence": "The document does
      ↪  not state that personal data
      ↪  will be processed unlawfully,
      ↪  unfairly, or without
      ↪  transparency, nor does it
      ↪  contradict principles of
      ↪  minimization, retention, or
      ↪  accountability."
    },
    {
      "regulation_id": "GDPR Article 3",
      "regulation_name": "GDPR
      ↪  Territorial scope",
```

```
        "contradiction_found": true,
        "compliance_status":
        ↪  "NON_COMPLIANT",
        "evidence": "Our Services are not
        ↪  intended for distribution to
        ↪  or use in any country or
        ↪  territory where such use
        ↪  would subject us to
        ↪  regulations in another
        ↪  country or territory."
      }
    ]
}
```

B Evaluator Prompt (LLM-as-a-Judge)

This section provides the prompt regarding the *LLM evaluation method* used for the model-based evaluator.

Evaluator Prompt

You are a legal-compliance evaluation assistant.

You will receive a JSON object representing the output of a GDPR contradiction-checking system. Your task is to assess whether the system's final conclusion is logically and legally justified by its own explanation and cited evidence.

This is not an adversarial task. Do not attempt to fabricate weaknesses. Respond with NO only if you can identify a clear and concrete flaw in the system's reasoning.

Evaluation procedure:

1. Identify the system's primary conclusion:
   - Determine whether it asserts that a direct contradiction exists or does not exist.

2. Follow the reasoning path:
   - GDPR rule or principle referenced
   - Claim made about the document
   - Supporting evidence (e.g., quotations or references)

3. Verify internal consistency:
   - Does the cited GDPR rule genuinely support the stated claim?
   - Does the evidence directly substantiate the claim, or is it irrelevant or misaligned?
   - Are there unjustified logical jumps (e.g., claims extending beyond what the evidence demonstrates)?

4. Conditions for answering YES:
   - The reasoning is coherent and the evidence reasonably supports the conclusion, even if the explanation is high-level or lacks fine-grained detail.
   - The system correctly interprets missing or vague information as the absence of a contradiction.

5. Conditions requiring NO (at least one must apply):
   - Misinterpretation of a GDPR rule or legal principle
   - Evidence that fails to support the stated legal claim
   - Logical inconsistency between the conclusion and the explanation
   - A conclusion that conflicts with the system's own evidence or raw_data

Final output rule:

You MUST return ONLY ONE WORD:
YES
or
NO

Meaning:

- YES = The system's conclusion is logically and legally supported by its own explanation and evidence.
- NO = A concrete logical or legal error has been identified in the system's reasoning.

Do not include any explanations, formatting, or additional text in your response.