MARIA: A Multi-Agent Regulatory Intelligence Architecture

Naomie Halioua[1], Alexandre Bloch[1], and Anaelle Guez[1]

[1]Cleo Labs, Paris, France

February 2026

Abstract

Determining which regulatory instruments apply to a given organization across jurisdictions and languages is a prerequisite for any compliance process, yet this upstream task has not been formalized as a computational problem. We call it *regulatory intelligence* and distinguish it from compliance checking, obligation extraction, and regulatory change detection, all of which assume the relevant regulation has already been identified.

We introduce MARIA (Multi-Agent Regulatory Intelligence Architecture), a system that automates this task. Given a company domain name, MARIA identifies applicable regulations across 19 geographic regions, 16 regulatory domains, and eight languages, then monitors regulatory developments, scores risk, and generates impact assessments.

We evaluate MARIA on seven companies spanning seven sectors (7,676 regulations total) using a hybrid protocol combining LLM-based evaluation at scale with targeted human review of 130 disputed classifications. The primary evaluator (GPT-5.2) estimates 73%–92% inclusive precision across the five companies where it renders informative verdicts. Human review reveals systematic evaluator conservatism: adjusted for identified bias, precision on binding regulations reaches 93%–97%, with zero fabricated entries. Errors are concentrated and sector-dependent: geographic scope mismatch, activity scope mismatch, and proposed-regulation rejection account for 87% of false positives.

Two methodological findings generalize beyond our system. First, regulatory applicability is not binary: directives, extraterritorial instruments, and role-dependent obligations require multi-valued classification that existing benchmarks do not support. Second, LLM-as-judge evaluation is unreliable for legal classification: two frontier models produce near-zero agreement ($\kappa \leq 0.340$) on the same task. **Keywords:** regulatory technology, regulatory intelligence, multi-agent AI, LLM evaluation, compliance automation

1 Introduction

Automated processing of regulatory texts has advanced considerably in recent years: compliance checking [14, 21], obligation extraction [26, 27], normative change modeling [13, 20, 25], and cross-lingual legislative retrieval [28] now benefit from mature computational methods. However, these approaches share a common assumption: *the relevant regulation has already been identified*. The upstream problem of determining, for a given organization, which regulatory instruments apply across jurisdictions and languages has received comparatively little attention in the literature.

We call this problem *regulatory intelligence* and argue that it constitutes a distinct computational task, formally different from compliance checking (which assumes a known rule), obligation extraction (which assumes a known text), and regulatory change detection (which assumes a known corpus). Our goal is to automate entity-specific regulatory mapping across business sectors, integrating compliance monitoring seamlessly into organizational workflows.

Scale and ubiquity of the problem

The practical importance of this task is growing rapidly. The EU alone has enacted over 13,000 binding legal acts since 2019 [1], and GDPR enforcement has exceeded €4.5 billion in cumulative fines [2]. Recent instruments including the AI Act [3], NIS2 Directive [4], Digital Markets Act [5], Data Act [6], and DORA [7] have each created new compliance obligations. In the US, 20 states have enacted comprehensive privacy legislation [8]. China's PIPL [9], India's DPDPA [11], and Japan's APPI [10] add further cross-jurisdictional complexity. A company operating across EU, US, and APAC markets may need to monitor 50+ regulatory frameworks simultaneously.

Several categories of commercial tools address aspects of this challenge. Compliance automation platforms (e.g., Norm AI, Secureframe) help apply known rules to business processes. Legal AI assistants (Harvey, CoCounsel) support research within existing corpora. GRC platforms (OneTrust, ServiceNow GRC) provide compliance workflow infrastructure. Regulatory change management services (Regology [38], Compliance.ai/Archer [39], CUBE [40]) track regulatory updates, typically for specific sectors. None of these platforms has published evaluation results on common benchmarks or disclosed systematic error analyses.

The formalization gap

The gap we address is therefore not one of industry capability but of *academic formalization*. No published work defines entity-specific regulatory mapping as a computational task, evaluates it with reproducible protocols, or provides error taxonomies that would enable systematic comparison. Legal NLP benchmarks such as LegalBench [23] and LEXTREME [22] evaluate comprehension of known texts; production systems such as STARA [34] and the tax-law work of Nay et al. [24] operate on fixed, curated corpora. Bajwa et al. [29] identify compliance monitoring as a critical open problem and confirm that performance degrades sharply outside English, a structural barrier for cross-jurisdictional coverage.

Contributions

We make three contributions. First, we formalize regulatory intelligence as a five-stage computational task with an explicit instrument ontology distinguishing *binding applicability* from *monitoring relevance* (Section 2). Second, we present MARIA, a system implementing this decomposition through orchestrated LLM specialization across 19 geographic regions, 16 regulatory domains, and eight languages (Section 3). Third, we evaluate MARIA on seven companies (7,676 regulations) using a hybrid human–LLM protocol that produces both a sector-dependent error taxonomy and systematic evidence of LLM-as-judge unreliability in legal classification (Sections 5–6).

2 Formalizing Regulatory Intelligence

2.1 Instrument Ontology: Binding Applicability vs. Monitoring Relevance

Regulatory intelligence must distinguish between two related but distinct questions:

- **Binding applicability:** Is this instrument legally enforceable on this entity today? This is a legal determination with a definite answer for a given point in time.
- **Monitoring relevance:** Should this entity's compliance team track this instrument? This is a practical determination that extends beyond binding force to include proposed legislation, voluntary standards that may become contractual requirements, and regulations applicable to business partners.

The distinction matters because regulatory instruments span a spectrum of binding force:

Instrument type	Binding?	Monitor?
Domestic regulation (e.g., GDPR)	Yes	Yes
EU directive (pre-transposition)	No*	Yes
Proposed legislation (e.g., PSD3)	No	Yes
Voluntary standard (e.g., ISO 27001)	No	Often
Foreign regulation (extraterritorial)	Depends	Yes
Enforcement guidance	No	Yes

Table 1: Instrument types and their binding/monitoring status. *Directives bind member states, not companies directly, but transposing legislation does.

MARIA is designed for *monitoring relevance*: it casts a deliberately wide net, including proposed regulations, voluntary standards with sectoral adoption, and instruments with extraterritorial reach. This is an explicit design choice—a compliance team that learns about PSD3 only after it enters into force has failed at regulatory intelligence. But it means that evaluating MARIA against a strict *binding applicability* standard will systematically overcount false positives: every proposed regulation, every voluntary standard, and every extraterritorial instrument with uncertain reach will appear as an error.

We formalize this by reporting two metrics throughout the evaluation: *strict precision* (measuring binding applicability: only "Yes" = TP) and *inclusive precision* (measuring monitoring relevance: "Yes" + "Partial" = TP, only "No" = FP). The gap between the two quantifies the system's coverage of the gray zone that compliance professionals consider valuable but that a strict legal test would exclude.

2.2 Regulatory Intelligence as a Distinct Task

We define *regulatory intelligence* as the task of determining, for a given organization, which regulatory frameworks apply, what has changed recently, and how urgent each change is. This definition distinguishes regulatory intelligence from three adjacent tasks that the literature has addressed:

- **Compliance checking** [14, 21] verifies whether a known process satisfies a known rule. It assumes both the rule and the process are given.
- **Obligation extraction** [26, 27] identifies what a given legal text requires. It assumes the text is given and relevant.
- **Regulatory change detection** [25, 28] identifies modifications within a known corpus. It assumes the corpus is defined.

Regulatory intelligence sits upstream of all three: it determines *which* rules, texts, and corpora are relevant in the first place, given only a company identifier. The distinction is not merely practical—it reflects a different computational structure. Compliance checking is a satisfaction problem (does process P satisfy rule R?). Obligation extraction is an information extraction problem (what does text T require?). Regulatory intelligence is an open-world classification problem: given an entity E, identify the set of applicable regulatory instruments $\mathcal{R}_E$ from an unbounded, multilingual, and continuously evolving regulatory landscape.

2.3 Task Decomposition

We decompose regulatory intelligence into five sequential stages (Figure 1):

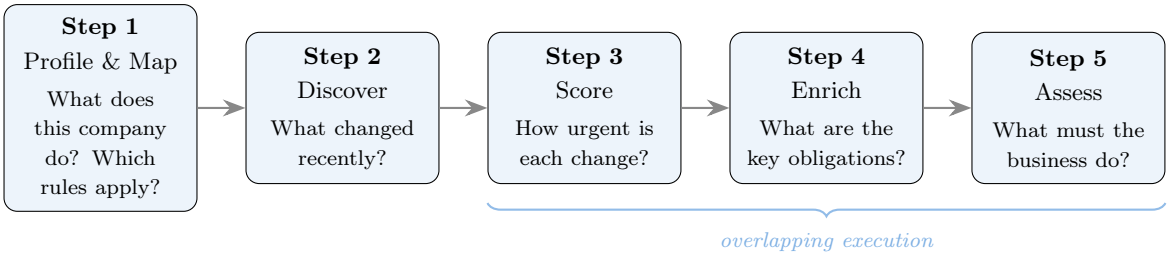

Figure 1: The MARIA five-stage regulatory intelligence pipeline.

We refer to this five-stage decomposition and its implementation as MARIA. The stages are logically sequential—each depends on the output of its predecessor—but the implementation overlaps execution where possible (Steps 3–5 process articles as they become available rather than in batch). Each stage addresses a qualitatively different reasoning challenge. We illustrate with a concrete example: determining whether a French AI company that serves US customers via a cloud API must comply with the California Consumer Privacy Act (CCPA).

Stage 1: Profile & Map. Given a company domain name, the system must determine what the company does, where it operates, what data it processes, and which sectors it belongs to. For our example, this requires inferring that a French AI company serving US customers processes personal data of California residents—a factual determination that requires entity understanding beyond what a corporate registry or keyword match provides. The system must then *map* this profile to applicable regulatory frameworks. The CCPA applies to entities that do business in California and meet revenue or data-volume thresholds—but the company's French headquarters does not exempt it, because "doing business in California" has been interpreted to include providing services to California residents via the internet. This mapping requires legal knowledge about extraterritorial reach that is jurisdiction-specific.

Stage 2: Discover. Having identified the CCPA as applicable, the system must detect *recent changes*: proposed amendments, enforcement actions, and regulatory guidance. The California Privacy Protection Agency issues guidance in English, but relevant commentary may appear in French legal blogs, EU adequacy discussions, or APAC data-transfer analyses. Monitoring must be multilingual and temporally continuous.

Stage 3: Score. Not all changes are equally urgent. A CPPA enforcement action against a similarly situated company is more urgent than a proposed amendment in committee. The system must assess relevance to the specific company (does this affect AI companies specifically?), likelihood of impact (is enforcement imminent?), and severity (what are the penalties?). This requires combining legal knowledge with entity-specific context.

Stage 4: Enrich. For high-scoring changes, the system extracts specific obligations: what must the company do, by when, and what are the consequences of non-compliance?

Stage 5: Assess. The system generates a company-specific impact assessment: given this company's data processing activities, what operational changes does this regulatory development require?

2.4 Why Regulatory Intelligence Requires AI

Assessing the relevance and applicability of regulations, as well as the specific compliance of company procedures, requires complex information processing and inference that is beyond any

form of information retrieval. Regulatory texts contain deontic expressions—obligations, permissions, prohibitions—whose scope depends on entity characteristics, jurisdictional context, and temporal validity. Traditional compliance approaches formalize these as deontic logic operators [14], but regulatory intelligence must perform similar reasoning implicitly across hundreds of instruments in multiple languages, without pre-formalized rule sets.

Consider the legal knowledge requirement. The same term—"data controller"—carries different legal obligations under the GDPR (EU), PIPL (China), and DPDPA (India). An EU Directive is not directly binding on companies; it must be transposed into national law, meaning the *same* directive produces 27 different national implementations. A proposed regulation that has passed committee vote but not yet entered into force is not binding today but may require compliance preparation now. These distinctions require reasoning about the nature and status of regulatory instruments—a form of implicit deontic reasoning that LLMs can approximate but that keyword search cannot perform.

This combination of capabilities—entity understanding, legal knowledge, temporal awareness, and domain reasoning—is why regulatory intelligence is an AI and Law problem in the sense of Prakken and Sartor [15] and Livermore and Rockmore [16]. AI techniques also have the potential to support explainability, allowing compliance teams to trace how a regulatory applicability determination was reached.

2.5 Relationship to the Computational Compliance Literature

Marino et al. [32] recently proposed a comprehensive blueprint for "computational compliance" spanning norm identification, interpretation, implementation, and monitoring. Our regulatory intelligence formalization maps to their first stage—norm identification—but with a critical difference: we start from the company rather than from a regulatory corpus. Videsjorden et al. [33] decomposed AI Act compliance into modular agent tasks, following a similar decomposition principle. We apply the same principle but address a different question: not "how do we comply with a known regulation?" but "which regulations apply in the first place?"

3 MARIA: System Design

We implement the five-stage decomposition through an architecture we call MARIA (Multi-Agent Regulatory Intelligence Architecture). MARIA follows an agentic design [17]: a centralized orchestrator delegates each subtask to a specialized LLM call with a task-specific prompt, context window, and computational budget. The orchestrator maintains all state and controls execution order; individual calls do not communicate with each other. This centralized architecture, which Jin et al. [35] showed to be more cost-effective than decentralized coordination for decomposable tasks, draws on ReAct [17] for reasoning-action loops and Tomasev et al. [19] for task delegation patterns.

3.1 Risk Scoring

Each finding is scored at two levels. First, six content dimensions (source authority, content depth, relevance, timeliness, actionability, and regulatory-act relevance) produce a quality score that determines whether a finding proceeds to enrichment. This filters noise: press releases mentioning a regulation in passing are scored lower than enforcement actions or legislative analyses. Second, two composite risk factors following ISO 31000—legal severity and legal likelihood, each on a five-point scale—produce a risk level:

$$\text{Risk Level} \;=\; \mathit{legal\_severity} \;\times\; \mathit{legal\_likelihood}\,, \qquad \text{each} \in \{1,\ldots,5\} \tag{1}$$

The risk level drives urgency classification (green, yellow, orange, red) surfaced to the compliance team. The two scoring layers serve different purposes: content scoring filters noise; risk scoring

prioritizes action. Both are computed by dedicated LLM calls with structured output schemas, ensuring that scores are consistent and auditable.

3.2 Search Infrastructure

A single MARIA execution requires substantial search volume. The upper bound decomposes as:

$$S_{\text{total}} \leq \underbrace{(20+\varepsilon)}_{\text{profiling}} + \underbrace{(C \times 4)}_{\text{mapping}} + 5 \cdot N_{\text{reg}}\,, \qquad \varepsilon \in [0,\, 28],\;\; C \leq 37 \tag{2}$$

where C is the number of regulatory chunks selected from a pool of 37 (19 geographic, 2 cross-domain, 16 sector-specific). A preliminary LLM call selects the relevant subset based on the company profile, so C varies per company. For a typical company with $C \approx 25$ and 50–200 mapped regulations, this yields roughly 370–1,200 web searches per run. MARIA issues queries in seven languages (French, German, Spanish, Portuguese, Chinese, Japanese, Korean) in addition to English, with language-specific query templates adapted to each jurisdiction's regulatory vocabulary.

3.3 Orchestration and Reliability

A single MARIA execution orchestrates over 200 specialized LLM calls: approximately 40–80 for company profiling and regulation mapping (including three parallel synthesis agents, cross-referencing, and per-chunk mapping), and one call per article for each of the scoring, enrichment, and assessment stages. Combined with the web searches described above, a typical run involves hundreds of parallel inferences.

Computational budgets are tiered by subtask: tasks requiring legal judgment (regulation mapping, impact assessment) receive higher reasoning budgets than structured extraction tasks (query planning, article parsing). This reduces cost substantially compared to a uniform maximum budget, with negligible quality degradation on simpler tasks.

Callback-driven streaming reduces end-to-end latency: each article flows through scoring, enrichment, and assessment as soon as it is ready, rather than waiting for all articles to complete a stage before the next begins. Each processed item is written to the database as it completes, enabling resume-on-retry after failures. This reliability architecture mitigates the risks identified by Bandara et al. [30] as the primary barrier to production deployment of agentic systems: in compliance monitoring, a silently crashed system is particularly dangerous because the compliance team believes regulatory changes are being tracked when they are not.

4 Previous and Related Work

Regulatory intelligence draws on and extends several research areas. We review each in turn, identifying what has been achieved and the aspects that remain unaddressed.

Computational legal reasoning. The tradition running from Bench-Capon and Sartor [12] through Prakken and Sartor [15] has formalized legal reasoning as argumentation, defeasible logic, and case-based reasoning. A central concern has been the analysis of deontic content—obligations, permissions, and prohibitions expressed in regulatory texts—whether through formal deontic logic or through detection of deontic expressions in natural language. Governatori et al. [14] and Hashmi et al. [21] developed methods for checking whether business processes comply with specified norms, relying on explicit formalization of regulatory requirements. Boella et al. [13] and Palmirani et al. [20] formalized how norms change over time, providing the theoretical basis for norm dynamics. More recently, Dal Pont et al. [26] and Sapienza and Palmirani [27]

extract obligations from legal texts using LLMs, which perform implicit deontic reasoning without requiring formal rule encoding. Corazza et al. [25] model reporting requests with hybrid AI. Zilli et al. [28] demonstrated that agentic approaches outperform static retrieval for cross-lingual legislative tasks.

These approaches share a common starting point: a *known* legal text or corpus. They answer "what does this regulation require?" or "does this process comply?" but not "which regulations apply to this company?" Our work addresses the identification stage that precedes their analyses. In the taxonomy of Marino et al. [32], we contribute to the norm identification stage, but with an entity-centric rather than corpus-centric formulation. The deontic reasoning that these works formalize explicitly is performed implicitly by the LLM calls in our pipeline, which must determine applicability without pre-encoded rule sets.

Natural language processing for legal texts. Legal NLP encompasses tasks ranging from named entity recognition in regulatory documents to classification of deontic expressions and extraction of normative content. Guha et al.'s LegalBench [23] demonstrated LLM competence on issue-spotting tasks but weaker performance on statutory interpretation. Niklaus et al.'s LEXTREME [22] showed that NLP performance on legal texts drops sharply outside English, affecting entity recognition, classification, and information extraction alike. Bajwa et al. [29] confirmed this multilingual degradation remains acute in 2025. Gui et al. [36] constructed an error taxonomy for LLM legal reasoning that informs our own error analysis.

For regulatory intelligence, the multilingual gap is a structural barrier rather than a secondary limitation. A system that monitors only English-language sources provides effectively no coverage for jurisdictions where regulatory texts, enforcement actions, and professional commentary are published in Chinese, Japanese, Korean, German, or Portuguese. Our system issues queries in seven non-English languages with jurisdiction-specific templates, treating multilingual coverage as a first-class design requirement.

Deployed legal AI systems validate the approach on bounded corpora. Surani and Ho's STARA [34] completes statutory surveys 2.7× more accurately than a base LLM at 86 cents per run. Nay et al. [24] showed frontier LLMs exhibit capability jumps on US tax law. Both demonstrate that LLMs can perform sophisticated legal reasoning, but on fixed, curated corpora where the relevant documents are known in advance.

Agentic systems for legal and regulatory text processing. The diversity of tasks required for regulatory text processing—profiling, mapping, searching, scoring, and assessing—and the need to coordinate them makes this problem well-suited for an agentic approach. Videsjorden et al. [33] decomposed AI Act compliance into modular agent tasks. We apply the same decomposition principle but address a different question: not "how do we comply with a known regulation?" but "which regulations apply?" The architecture we adopt—centralized orchestration rather than peer-to-peer dialogue—serves as mitigation for the reliability risks that Bandara et al. [30] and Langfuse [37] have documented: in agentic deployments, failure modes propagate unpredictably, and a centralized controller simplifies recovery and state management.

The industrial RegTech landscape is active but academically opaque. Several commercial platforms address aspects of the regulatory intelligence task: Regology [38] offers entity-specific regulation mapping with AI-assisted change monitoring; Compliance.ai (now Archer) [39] provides regulatory change feeds primarily for financial services; CUBE [40] automates regulatory intelligence with curated content. These platforms demonstrate market demand and technical feasibility. However, none has published evaluation results on common benchmarks, disclosed error rates, or characterized failure modes systematically. Our contribution is not that these

systems do not exist, but that the *academic infrastructure* for evaluating them—task formalization, evaluation protocols, error taxonomies, and public benchmarks—does not. MARIA serves as the experimental vehicle for building this infrastructure.

5 Evaluation

We evaluate MARIA on the regulation mapping stage: given a company, does the system correctly identify which regulations apply? We focus on *precision*—the fraction of identified regulations that are genuinely applicable—as the primary metric, because for a system designed to cast a wide net (§2.1), the most actionable question is how much noise the compliance team must filter. Recall (what the system misses) is discussed qualitatively in the limitations; we do not claim to have established a recall floor.

In the absence of accepted benchmarks for multi-jurisdictional regulatory applicability, we have designed a hybrid evaluation method combining LLM-based classification at scale with targeted human review. At 5–10 minutes per regulation, full human verification of 7,676 regulations would require over 800 analyst hours. Our protocol uses LLM judges for coverage, then audits their verdicts with human review on disputed cases. This introduces a complication we take seriously: the LLM judges are themselves unreliable, as we will show. The precision figures we report are therefore *LLM-estimated precision*: proxy measurements whose biases we characterize rather than ground truth we claim to have established.

5.1 Setup

To ensure that our evaluation covers a wide range of regulatory problems, we have selected companies from diverse industrial sectors that share some regulatory requirements (e.g., data protection) but also exhibit sector-specific compliance challenges (e.g., financial regulation for Revolut, content moderation for TikTok, advertising law for Havas). We select seven companies spanning seven sectors and four headquarters countries (Table 2). The sample is designed to vary along three dimensions that we expect to affect both system performance and evaluator behavior: sector (from physical infrastructure to social media), geographic scope (France-only to 180+ markets), and headquarters location (France, US, UK, China/Ireland, Sweden).

#	Company	Sector	HQ	Scope	Complexity
1	Vinci Autoroutes	Infrastructure / IoT	FR	France + limited global	Medium
2	Mistral AI	AI / LLM	FR	France + global API	Medium
3	Havas	Advertising / Media	FR	France + 100+ countries	Very high
4	Deel	HR SaaS / EOR	US	US + EU + 150+ countries	High
5	Revolut	Fintech	UK	UK + EU + global	High
6	TikTok	Social Media	CN/IE	Global (150+ countries)	Very high
7	Spotify	Music / Podcast Streaming	SE	Global (180+ markets)	High

Table 2: Evaluation sample: seven companies across four headquarters countries, seven sectors, and four complexity levels.

The evaluation proceeds in two stages with distinct roles. The *primary evaluation* uses two LLM models—GPT-5.2 (OpenAI) and Gemini (Google)—to classify every regulation as Yes (directly applicable), Partial (applicable with caveats), or No (false positive). Each model receives the company profile but not the system's relevance scores. This automated stage provides coverage across all 7,676 regulations. The *human audit* then reviews regulations where the LLM judges disagree or assign non-Yes verdicts, targeting the cases most likely to reveal systematic evaluator bias. A legal analyst verified 130 regulations across the seven companies against primary legal sources. This design maximizes information about evaluator bias per unit of analyst

time, at the cost of representativeness: the human sample is biased toward disagreement points.

We note that the term "LLM-as-judge" typically refers to LLMs evaluating other LLMs' generated text. Our usage is different: here, GPT-5.2 and Gemini evaluate the *correctness of a classification* (is this regulation applicable to this company?), a task closer to legal analysis than to output quality assessment.

We report *inclusive precision* as the primary metric: Yes + Partial = true positive, only No = false positive. This aligns with the instrument ontology (§2.1): inclusive precision approximates *monitoring relevance* (is this regulation worth tracking?), while strict precision (only Yes = TP) approximates *binding applicability* (is it enforceable today?).

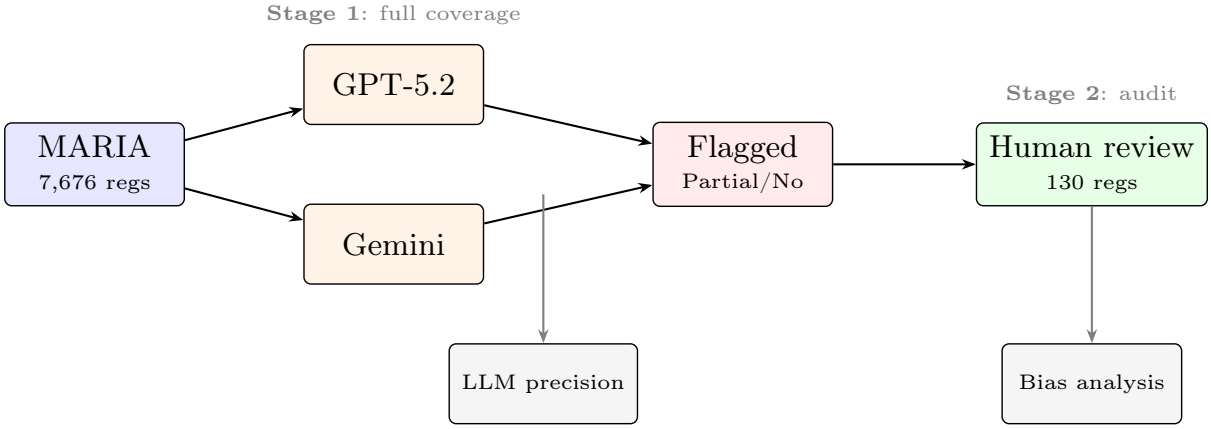

Figure 2: Evaluation protocol. Stage 1: two LLM judges classify all 7,676 regulations. Stage 2: legal analyst reviews 130 flagged cases for bias characterization.

5.2 Precision and Error Structure

Table 3 presents the central result. Across five companies where GPT-5.2 renders informative verdicts, inclusive precision ranges from 72.6% to 91.6%. Two companies—Deel (99.9%) and TikTok (100.0%)—sit outside this range, and we will argue their scores reflect evaluator saturation rather than system accuracy.

Company	Sector	HQ	Regs	TP	FP	Incl.
Vinci Autoroutes	Infrastructure	FR	500	458	42	91.6%
Mistral AI	AI / LLM	FR	1,276	1,024	252	80.3%
Havas	Advertising	FR	1,100	799	301	72.6%
Deel	HR SaaS / EOR	US	1,200	1,199	1	99.9%†
Revolut	Fintech	UK	1,200	1,088	112	90.7%
TikTok	Social Media	CN/IE	1,200	1,200	0	100.0%‡
Spotify	Streaming	SE	1,200	1,057	143	88.1%
Overall			**7,676**	**6,825**	**851**	**88.9%**

Table 3: Regulation mapping precision (GPT-5.2 as judge). These are *LLM-estimated* precision figures, not ground-truth measurements; see §5.3 for human calibration. †Deel's 99.9% reflects 94% Partial classification rather than confident identification; see §5.3. ‡TikTok's 100% reflects zero FPs but 47.9% Partial, indicating evaluator saturation; see §5.3.

The numbers separate into three regimes. Companies with geographically focused operations achieve the highest precision: Vinci (91.6%, primarily French infrastructure) and Revolut (90.7%, heavily regulated fintech). Globally distributed companies fall lower: Spotify (88.1%), Mistral (80.3%), and Havas (72.6%, the worst in the study). The third regime—Deel and TikTok—is not

really about system quality at all. Deel's 94% Partial rate means GPT-5.2 almost never commits to a verdict; TikTok's zero rejections means it cannot distinguish applicable from inapplicable regulations for a platform present in 150+ countries. For these two companies, the Partial rate is the real signal, not precision.

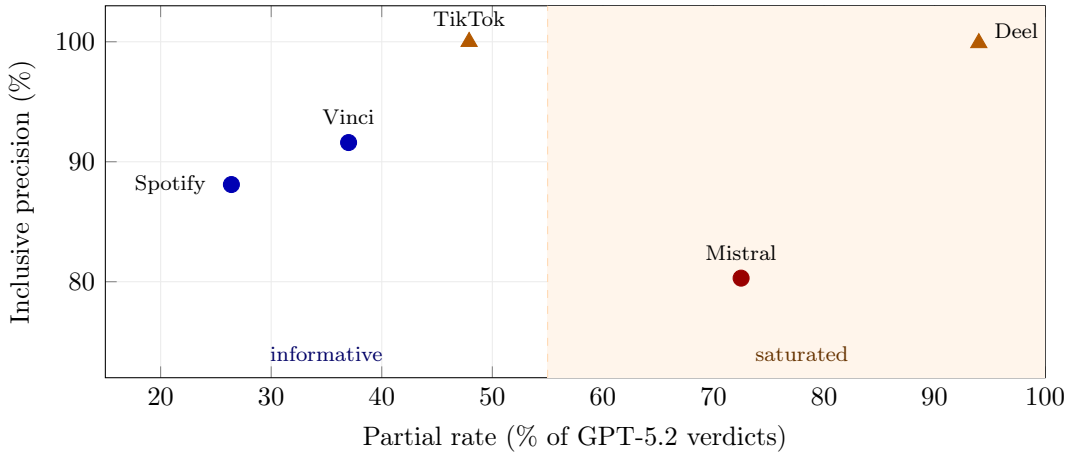

Figure 3: Evaluator informativeness. Shaded: high Partial rate = GPT-5.2 hedges. Deel/TikTok: near-perfect because evaluator cannot reject. Havas/Revolut omitted.

This pattern holds at the regional level (Table 4). French and EU regulations achieve >90% precision across all companies. Precision degrades for non-EU jurisdictions, with the magnitude depending on whether the evaluator correctly accounts for the company's local presence. The most striking failure: GPT-5.2 achieves only **56.7% on the UK for Revolut**—a UK-headquartered, FCA-regulated fintech.

	Vinci		**Mistral**		**Havas**		**Deel**		**Revolut**		**Spotify**	
Region	n	**L.**	n	**L.**	n	**L.**	n	**L.**	n	**L.**	n	**L.**
France	178	98.3	64	90.6	55	96.4	57	100†	65	96.9	64	89.1
Sweden	—	—	—	—	—	—	—	—	—	—	23	91.3
EU (s. + MS)	151	98.7	542	86.5	284	94.4	579	99.8†	518	97.1	378	88.4
US	—	—	186	80.1	118	79.7	144	100†	101	86.1	168	86.9
UK	—	—	68	66.2	46	54.3	47	100†	90	**56.7**‡	48	91.7
APAC	—	—	236	69.1	294	56.5	146	100†	139	87.1	188	83.5
Lat. Am.	—	—	63	88.9	86	61.6	85	100†	29	82.8	56	96.4
Intl.	171	75.4	20	40.0	2	100	7	100†	72	79.2	86	83.7

Table 4: Regional inclusive precision (%, GPT-5.2). TikTok omitted: 100% across all regions (zero FPs). †Deel: uniform high precision reflects 94% Partial classification (§5.3). ‡Revolut: 56.7% on the UK—GPT-5.2 rejects UK regulations for a UK-headquartered company (§5.3).

Where does the system go wrong? The 851 false positives (Table 5, Figure 4) are not randomly distributed—they cluster by company and by error type. Three types account for 87% of all false positives: geographic scope mismatch (61.2%), proposed-regulation rejection (13.0%), and activity scope mismatch (12.5%). Each type is concentrated in specific sectors, which means each is addressable with a targeted fix.

A note on fabricated entries: we verify that every regulation MARIA identifies corresponds to a real legal instrument by checking the regulation name and number against official legal databases (EUR-Lex, national legislative portals, and international treaty repositories). Across all 7,676 regulations, zero fabricated entries were found—every identified regulation exists in a verifiable legal source, even when it is not applicable to the target company.

Error type	Vi.	Mi.	Ha.	De.	Re.	Ti.	Sp.	Total
Geographic scope	3	188	252	—	78	—	—	521 (61.2%)
Activity scope	—	—	—	—	—	—	106	106 (12.5%)
Proposed / not-in-force	—	—	47	—	34	—	30	111 (13.0%)
Non-binding instr.	—	55	—	—	—	—	—	55 (6.5%)
Voluntary standard	34	—	—	—	—	—	—	34 (4.0%)
Entity scope	3	9	2	1	—	—	7	22 (2.6%)
Technical standard	2	—	—	—	—	—	—	2 (0.2%)
Total FP	**42**	**252**	**301**	**1**	**112**	**0**	**143**	**851**

Table 5: Error taxonomy ($N_{\text{FP}} = 851$). Geographic scope mismatch dominates (61.2%). Each company's errors are dominated by a single type. Deel's single FP understates its evaluation challenge—see Table 6.

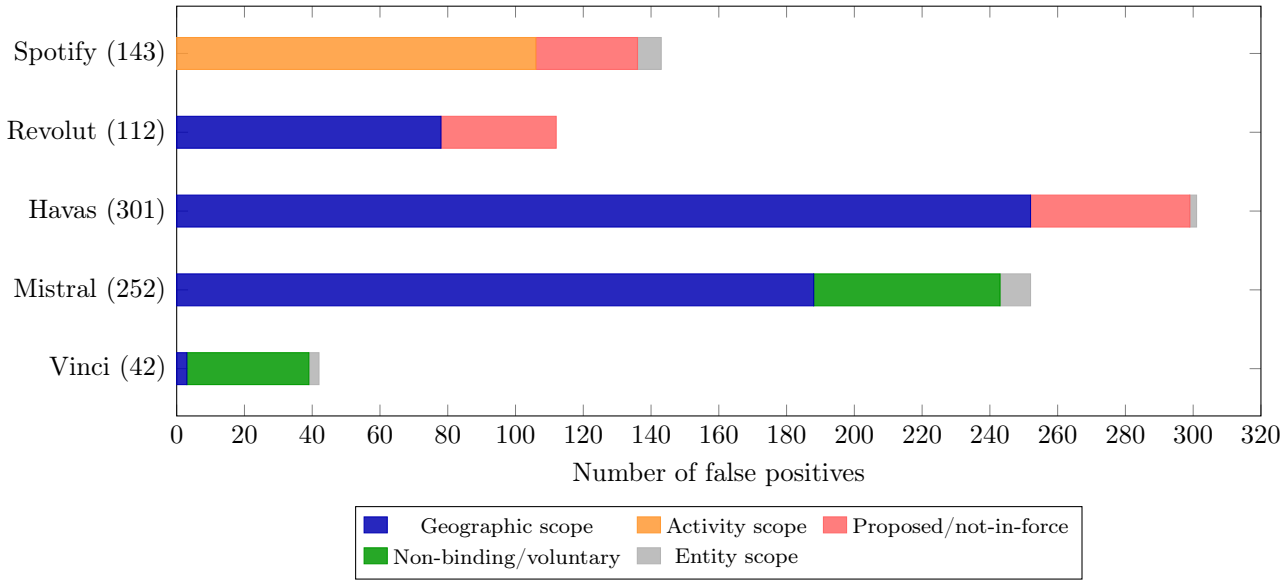

Figure 4: Error composition by company ($N_{\text{FP}} = 851$; Deel and TikTok omitted: 1 and 0 FP respectively). Each company's errors are dominated by a single type, making them amenable to targeted correction.

The pattern is striking: Vinci's errors are almost entirely voluntary standards (ISO 14001, GRI, TCFD) misclassified as binding regulations. Mistral and Havas are dominated by geographic overreach—the same error type, but with an important distinction. For Mistral, the system includes jurisdictions where a globally accessible API may or may not constitute "offering services"; the legal ambiguity is genuine. For Havas, GPT-5.2 rejects non-EU regulations despite confirmed local offices—this is evaluator failure, not system error. Spotify introduces a novel pattern: activity scope mismatch, where GPT-5.2 applies a narrow "music streaming" frame and rejects regulations related to podcast hosting, advertising, and content moderation. Revolut combines geographic mismatch (including 30 UK regulations rejected for a UK company) with proposed-regulation rejection.

Deel requires separate treatment. Its 99.9% precision and single false positive mask a deeper issue: GPT-5.2 classifies 94% of Deel's regulations as "Partial" (Table 6), with 83.3% citing the same justification—applicability depends on whether Deel acts as employer-of-record, payroll provider, or SaaS platform. This is not evaluator failure. It reflects a genuine property of Deel's business model: regulatory applicability varies by operational mode per jurisdiction.

Partial justification category	n	**%**
EOR / employer role dependency	940	83.3%
Cybersecurity scope (NIS2, etc.)	144	12.8%
Tax / VAT scope	22	2.0%
AI regulation scope (AI Act)	14	1.2%
Data protection scope	5	0.4%
Other conditional	3	0.3%
Total Partial	**1,128**	**100%**

Table 6: Deel: GPT-5.2 "Partial" justification categories (N = 1,128). 83.3% cite role ambiguity: applicability depends on Deel's operational mode (EOR vs. SaaS vs. payroll) in each jurisdiction.

MARIA's internal relevance scoring (critical, high, medium, low) shows weak but significant rank correlation with GPT-5.2 verdicts across all companies (Spearman $\rho = 0.118$–0.307, all $p < 0.001$). The more informative finding is that GPT-5.2 confidence discriminates strongly across verdict categories (Kruskal–Wallis $p < 0.001$ for every company, followed by Dunn's post-hoc test with Holm correction): the evaluator is systematically least confident on Partial verdicts and most confident on clear approvals and rejections, consistent with the interpretation that Partial captures genuine legal ambiguity.

5.3 Auditing the Judge

The precision figures reported above are only as reliable as the LLM judge that produced them. This section asks: how reliable is it? The answer—not very—turns out to be the most interesting finding of the evaluation.

Two models, near-zero agreement. We run Gemini as a second judge on all seven companies (Table 7). Each evaluation uses a single LLM call per regulation with identical prompt templates, temperature 0, and the full company profile in context. No retrieval augmentation is used; both models operate from the regulation name and company profile alone. Under these conditions, the results are striking. Gemini approves >96% of regulations for Vinci and Havas, rejects 53% for Mistral, then returns to high-approval patterns for Deel, TikTok, and Spotify.

Cohen's κ never exceeds 0.340, indicating agreement no better than chance for most company pairs.

	Gemini verdict distribution				
Company	**Yes**	**Partial**	**No**	**Cohen's κ**	**Spearman ρ**
Vinci Autoroutes	99.6%	0.4%	0%	0.008	—
Havas	96.1%	3.9%	0%	−0.018	0.281
Mistral AI	19.4%	27.6%	53.0%	0.234	0.469
Deel	75.9%	24.1%	0%	0.029	0.113
Revolut	53.6%	33.8%	12.7%	N/A*	N/A*
TikTok	40.7%	59.3%	0%	0.340	0.349
Spotify	74.3%	25.7%	0%	0.230	0.388

Table 7: Cross-model evaluator reliability. Gemini's behavior is company-dependent and unpredictable. *Revolut's GPT and Gemini evaluations were conducted on different pipeline runs; cross-model statistics cannot be computed.

Several hypotheses could explain Gemini's inconsistency. The behavior may reflect sensitivity to prompt length (company profiles vary from 500 to 3,000 tokens), differential training coverage across jurisdictions, or an interaction between the model's context window handling and the volume of regulations evaluated per session. We cannot rule out that a different prompting strategy—for instance, providing structured company profiles or using retrieval-augmented evaluation—would yield more stable Gemini results. However, the fact that GPT-5.2 produces consistent directional bias under identical conditions suggests that the instability is at least partly model-specific rather than a pure artifact of our evaluation setup.

We use GPT-5.2 as the primary judge because its conservatism produces a consistent directional bias, making reported precision a **lower bound**. But Gemini's erratic behavior, whether intrinsic or usage-dependent, reinforces a broader concern: single-model evaluation of legal classification tasks cannot be assumed reliable without cross-validation.

What the human sees differently. Legal analysts reviewed 130 regulations flagged by at least one LLM judge (Table 8). The sample is biased toward disagreement by design, so the divergence rates overstate population-level disagreement. But the *patterns* of divergence are systematic and reveal four distinct failure modes of the LLM evaluator.

Company	n	**H–GPT**	**H–Gem.**	**Key divergence**
Vinci	46	10.9%	—	EU directives: Human Yes → GPT Partial
Mistral	21	38.1%	—	Non-domestic binding: Human Yes → GPT Partial
Havas	12	0.0%	83.3%	All non-EU: Human Yes → GPT No
Deel	16	50.0%	—	Bidirectional: GPT both stricter and more permissive
Revolut	13	7.7%	—	Proposed regs + UK regs: Human Yes → GPT No
TikTok	12	16.7%	—	Non-EU binding: Human Yes → GPT Partial
Spotify	10	40.0%	—	Activity scope: Human Yes → GPT No
Combined	**130**	—	—	GPT-5.2 systematically stricter (except Deel)

Table 8: Human–LLM calibration across 130 regulations. H–GPT = Human–GPT-5.2 disagreement rate. GPT-5.2 is systematically more conservative than human analysts for all companies except Deel.

Transposition and extraterritoriality (Vinci, Mistral, TikTok). The human classifies EU directives and non-domestic binding regulations as "Yes"; GPT-5.2 classifies them as "Partial," treating transposition and extraterritorial reach as conditions rather than established legal mechanisms. Both positions are defensible, but the gap is systematic: GPT-5.2 is one category stricter on all enacted non-domestic regulations.

Geographic frame anchoring (Havas, Revolut). GPT-5.2 applies a "French/EU company" frame regardless of actual operations. For Havas, it rejects all 12 reviewed non-EU regulations despite confirmed local offices—**0% agreement**. For Revolut, it rejects 30 UK regulations for a UK-headquartered company, with justifications reading "outside France/EU scope." This appears to be a usage-related issue rather than an intrinsic model limitation: the evaluator anchored on a geographic frame rather than incorporating the company profile. Providing a more structured company profile (e.g., an explicit list of countries with confirmed operations) would likely mitigate this pattern, though we did not test this hypothesis in the current evaluation.

Temporal scope disagreement (Revolut, Spotify). GPT-5.2 rejects PSD3, the FCA Safeguarding Regime, and Canada's AIDA as "not yet in force." The human analyst classifies them as applicable because compliance preparation is already required. This is a fundamental disagreement about what "applicable" means: the evaluator asks "is this binding today?"; compliance professionals ask "does this require action within my planning horizon?" Human–GPT agreement on Revolut: **7.7%**—the lowest in the study.

Activity scope and role ambiguity (Spotify, Deel). GPT-5.2 applies a reductive model of each company's activities. For Spotify, it rejects the EU DSA and Brazil's Marco Civil—applicable to a platform hosting user-generated podcast content—because it sees only a "music streaming" service. For Deel, disagreement is bidirectional: GPT is too strict on 5 of 16 reviewed regulations and too permissive on 3, depending on which operational mode it assumes.

Regulatory applicability is not binary. Across all evaluations, the "Partial" middle category absorbs 26–94% of GPT-5.2 verdicts depending on the company. Revolut adds a temporal dimension: 34 regulations classified "No" because they are proposed, yet treated as applicable by compliance professionals. The disagreement between human, GPT-5.2, and Gemini is not noise; it reflects genuine ambiguity in what "applicable" means for different instrument types.

We hypothesize that three factors drive this ambiguity. First, *legal instrument type*: EU directives, which bind member states but not companies directly, occupy a middle ground that human analysts resolve through practical transposition knowledge but that LLMs treat as conditional. Second, *temporal scope*: compliance professionals evaluate applicability over a planning horizon (6–18 months), while LLMs appear to evaluate current binding force. Third, *entity complexity*: multi-service companies (Deel, Spotify) create role-dependent or activity-dependent applicability that neither binary nor ternary schemes capture well.

The pattern across instrument types is systematic: for binding domestic regulations, all evaluators broadly agree; for EU directives and non-domestic instruments, GPT-5.2 downgrades to Partial where humans say Yes; for proposed regulations, GPT-5.2 rejects what humans consider monitoring-relevant; for voluntary standards, humans assign Partial while GPT-5.2 assigns No. Gemini's classifications are inconsistent across companies for every instrument type. The implications for evaluation methodology are discussed in Section 6.

Bounding the true precision. The four divergence patterns allow us to bound the true precision (Figure 5). Havas shows the largest correction (+22.9 pp), driven entirely by geographic frame anchoring. Mistral has no adjustment: its geographic errors reflect genuine legal ambiguity, not evaluator failure.

6 Discussion

The seven evaluations—Vinci Autoroutes (infrastructure), Mistral AI (AI/LLM), Havas (advertising), Deel (HR SaaS/EOR), Revolut (fintech), TikTok (social media), and Spotify (streaming)—span 7,676 regulations. We organize our findings around the three themes identified in the abstract—system precision and error structure, LLM-as-judge reliability, and the non-binary nature of regulatory applicability—decomposed into seven specific observations.

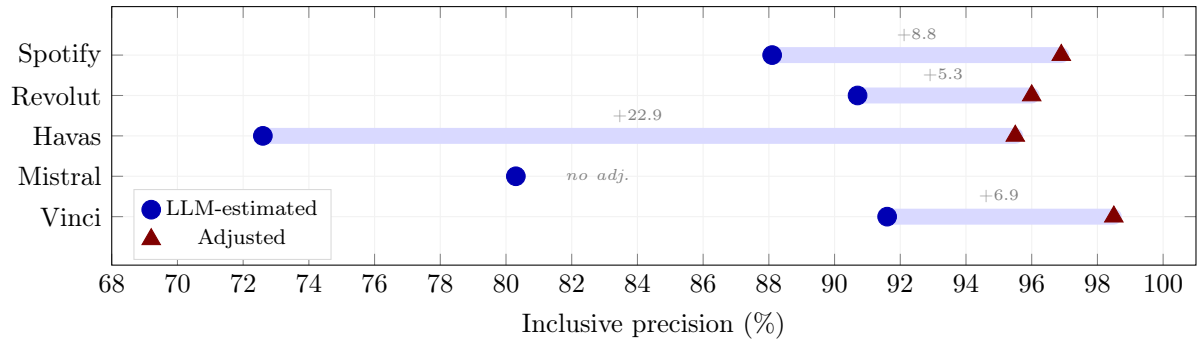

Figure 5: Precision bounds. Gap = evaluator bias. Havas: +22.9 pp. Deel/TikTok omitted (saturated).

1. High precision on binding regulations, with concentrated and company-type-dependent errors. On binding domestic and EU regulations, MARIA achieves >90% LLM-estimated inclusive precision across all companies except Havas (where evaluator failure, not system failure, drives the low score). The 851 false positives are not randomly distributed: 61.2% stem from geographic scope mismatch, 13.0% from proposed/not-yet-in-force rejection, and 12.5% from activity scope mismatch. These errors fall into two categories. Some are *structural* and will require permanent post-processing: voluntary standard inclusion (the system cannot distinguish binding regulations from industry standards without a binding/non-binding classifier) and proposed-regulation inclusion (a deliberate design choice for monitoring relevance that will always inflate precision under a strict applicability standard). Others are *addressable through system improvement*: geographic scope mismatch can be reduced by integrating verified operational footprint data from Step 1 into a geographic filter, and activity scope mismatch can be reduced by providing downstream stages with a structured decomposition of the company's service lines. Error mitigation must be company-type-aware, as different company profiles trigger fundamentally different error patterns.

2. LLM evaluator conservatism inflates false positive counts—but its severity depends on company context. Human calibration on 130 regulations reveals three GPT-5.2 behaviors. For companies with well-defined scopes (Vinci, Mistral, Havas), GPT-5.2 is systematically one category stricter. For companies with ambiguous operational modes (Deel) or broad global presence (TikTok), it collapses into "Partial" as a default. For non-French-headquartered companies (Revolut), it can *misidentify headquarters*, rejecting home-jurisdiction regulations. The reported precision figures are therefore **lower bounds**, but the magnitude of underestimation ranges from modest (systematic conservatism) to severe (headquarters misidentification).

3. Single-model evaluation is insufficient for legal classification tasks. Legal classification inherently involves underspecified categories ("applicable" admits multiple defensible interpretations), and different models resolve this underspecification differently. Gemini's behavior reverses between companies (99.6% approval for Vinci to 53.0% rejection for Mistral, back to 74.3% approval for Spotify), with Cohen's κ at or below 0.340. We hypothesize that this reflects differential sensitivity to entity complexity: Gemini may default to approval when the regulatory landscape is unfamiliar (Vinci, Havas) and to rejection when it has stronger training signal for the domain (AI regulation for Mistral). Alternatively, the instability may reflect context-length effects, as company profiles and regulation lists vary substantially in size. Regardless of the mechanism, evaluations of legal AI systems should report results from at least two independent models to expose such model-specific biases.

4. Regulatory applicability is not binary, and aggregate precision must be reported with verdict distributions. The three-category finding is reinforced across all seven evaluations. EU directives require transposition; voluntary standards are monitoring-relevant without being binding (§2.1); extraterritorial reach creates genuine ambiguity; multi-service business models create role-dependent applicability; proposed-but-imminent regulations occupy a gray zone between binding applicability and monitoring relevance; and activity-scope-dependent regulations are applicable only when the company's full service portfolio is recognized. A high inclusive precision with a low Partial rate (Vinci: 91.6%, 37% Partial; Spotify: 88.1%, 26.4% Partial) is more informative than a high inclusive precision with a high Partial rate (Deel: 99.9%, 94% Partial; TikTok: 100%, 47.9% Partial). Future benchmarks should adopt a multi-valued scheme and report full verdict distributions.

5. Compliance professionals and LLM evaluators disagree on temporal scope. Revolut's calibration reveals a fundamental disagreement: GPT-5.2 evaluates "is this binding today?" (binding applicability); compliance professionals evaluate "will this require action within my planning window?" (monitoring relevance, §2.1). This temporal disagreement inflates false positive counts by 34 regulations for Revolut and 30 for Spotify. Regulatory intelligence systems—which exist specifically to provide early warning—should adopt the compliance professional's temporal frame.

6. Evaluator discrimination collapses for maximally complex companies. TikTok's 100% inclusive precision with zero false positives demonstrates that for companies with sufficiently broad global operations, LLM evaluators lose the ability to reject regulations. Inclusive precision becomes uninformative; the Yes/Partial ratio and strict precision become the primary quality indicators. This has implications for benchmark design: evaluation metrics must account for evaluator saturation at high company complexity. The phenomenon is not limited to TikTok—Deel's 99.9% with 94% Partial shows a milder form of the same effect driven by business model ambiguity rather than geographic breadth.

7. Activity scope mismatch is a distinct and addressable error type. Spotify demonstrates that LLM evaluators can apply a reductive model of a company's service portfolio, rejecting regulations applicable to secondary business activities. This is distinct from geographic scope mismatch (wrong location), role ambiguity (unclear operational mode), and entity scope mismatch (wrong entity type): it reflects the evaluator's failure to map the company's full regulatory surface. The pattern is addressable: providing the evaluator with a structured decomposition of the company's service lines would likely reduce the 106 activity-scope FPs.

6.1 Implications for AI and Law Research

Beyond the seven empirical findings, this work raises three questions for the AI and Law community. MARIA's architecture—centralized orchestration of specialized LLM calls with tiered computational budgets—demonstrates that regulatory intelligence is tractable as an automated task when decomposed into stages with appropriate reasoning budgets.

First, **the absence of a regulatory applicability benchmark is a structural gap.** Existing legal NLP benchmarks (LegalBench, LEXTREME) evaluate comprehension and interpretation of known texts. No benchmark evaluates the upstream task of determining which regulations apply to a given entity across jurisdictions. Constructing such a benchmark—which would require ground-truth regulation portfolios for diverse company profiles, maintained across regulatory changes—is a significant undertaking, but our evaluation demonstrates both its feasibility and its necessity. The error taxonomy we present (geographic scope, activity scope,

voluntary standards, role ambiguity, proposed-regulation status, headquarters misidentification, evaluator saturation) could serve as a starting point for benchmark design.

Second, **the multi-valued nature of regulatory applicability challenges evaluation methodology.** The AI and Law community has long recognized that legal reasoning involves degrees of applicability, defeasibility, and context-dependence [12, 15]. Our empirical results quantify this for regulatory classification: binary Yes/No evaluation produces misleading precision figures, and the magnitude of the distortion is company-type-dependent. The role-ambiguity pattern (Deel), the temporal-scope disagreement (Revolut), and the activity-scope pattern (Spotify) suggest that regulatory applicability may require at minimum a five-valued scheme: *directly binding*, *conditionally applicable* (via transposition, extraterritorial reach, or operational mode), *activity-dependent* (applicable to specific service lines), *prospectively applicable* (proposed but imminent), and *not applicable*. This aligns with the norm dynamics work of Boella et al. [13] and Palmirani et al. [20], which formalized how norms transition between states—but extends it to the classification task.

Third, **LLM-as-judge reliability in legal domains requires systematic study.** Our cross-model analysis shows that two frontier LLMs produce near-zero agreement on regulatory classification. The finding that GPT-5.2 can misidentify a company's headquarters—applying a geographic frame from previous evaluations rather than from the target company's profile—suggests that LLM evaluators may exhibit a form of *anchoring bias* in sequential evaluation tasks. The finding that evaluator discrimination collapses entirely for TikTok suggests a form of *evaluator saturation* at high complexity. Gui et al.'s [36] error taxonomy for LLM legal reasoning does not yet cover evaluator-specific failure modes; our findings suggest this is a productive direction.

Limitations. This evaluation has several important limitations. First, the human calibration uses a **single annotator** per company; inter-annotator agreement is not measured. This means we cannot distinguish genuine human–LLM disagreement from annotator subjectivity. Future work should use dual annotation with adjudication. Second, the calibration sample (130 regulations) is **stratified toward disagreement points** rather than randomly sampled, which may overstate human–LLM divergence. Third, the evaluation relies on LLM judges whose unreliability we document: the precision figures are proxy measurements, not ground truth. Fourth, we provide **no baselines**—no comparison with keyword-based approaches, retrieval-only systems, monolithic LLM prompting, or commercial RegTech platforms. Without baselines, the contribution of MARIA's multi-stage architecture over simpler approaches remains undemonstrated. Fifth, the evaluation covers seven companies: three French-headquartered, one US, one UK, one Chinese/Irish, and one Swedish. Extension to additional headquarters countries (e.g., Germany, Israel) would test whether the evaluator's geographic bias generalizes. Sixth, Revolut's Gemini evaluation was conducted on a different pipeline run, precluding cross-model agreement statistics. Seventh, downstream pipeline stages (monitoring, scoring, impact assessment) are not yet evaluated. Eighth, recall remains unquantified: we cannot assess what regulations the system misses. Ninth, **reproducibility is limited**: the system's prompts, deduplication rules, and filtering logic are not published. We plan to release evaluation data and prompt templates in a companion repository.

7 Ethical Considerations

MARIA is designed to *augment* compliance professionals, not replace them: all outputs are scored recommendations requiring human review. Over-reliance could create a false sense of coverage. We mitigate this by surfacing confidence indicators and communicating source coverage limitations. The risk scoring model embeds normative judgments that are documented and configurable. The multilingual search component processes publicly available content only.

8 Conclusion

Regulatory intelligence—determining which regulations apply to a specific company, what changed, and how urgent it is—is a distinct computational problem that sits upstream of compliance checking, obligation extraction, and regulatory change detection. Despite its practical importance and the existence of commercial platforms addressing aspects of the problem [38–40], the academic literature has not formalized it as a task, and no published evaluation compares systems on common benchmarks. We have formalized this problem as a five-stage decomposition with an explicit instrument ontology distinguishing binding applicability from monitoring relevance (§2.1), and presented MARIA, an architecture implementing it through orchestrated LLM specialization across 19 geographic regions, 16 regulatory domains, and eight languages.

Evaluation across seven companies (7,676 regulations) demonstrates that MARIA achieves LLM-estimated inclusive precision of 73%–92% on five companies where the evaluator renders informative verdicts. Two further companies—Deel (99.9%) and TikTok (100.0%)—achieve near-perfect inclusive precision reflecting evaluator indecision rather than system accuracy. Human calibration on 130 regulations reveals that the LLM evaluator is systematically stricter than compliance professionals: these figures are lower bounds, while upper-bound adjusted estimates (assuming identified systematic evaluator errors are reclassified) range from 93% to 97%. Error types are sector-dependent and concentrated: geographic scope mismatch, activity scope mismatch, and proposed-regulation rejection account for 87% of all false positives.

The evaluation yields seven methodological findings with implications beyond our system. First, regulatory applicability requires a multi-valued classification aligned with the instrument ontology. Second, cross-model evaluation is essential: a second evaluator's behavior reversed entirely between companies. Third, LLM evaluators are systematically more conservative than human analysts, but the severity depends on company context. Fourth, aggregate precision must be reported alongside full verdict distributions. Fifth, compliance professionals and LLM evaluators disagree on temporal scope. Sixth, evaluator discrimination collapses for maximally complex companies. Seventh, activity scope mismatch is a distinct and addressable error pattern.

These findings, more than the system itself, constitute the primary scientific contribution. They point to a productive research agenda. First, future benchmarks for regulatory applicability should adopt a multi-valued classification scheme: *directly binding*, *applicable via transposition or soft-law*, *proposed but imminent*, and *not applicable*—with possible extensions for role-dependent and activity-scope-dependent applicability. Binary evaluation inflates false positive counts for the middle categories and masks systematic evaluator disagreement. Second, establishing baselines for entity-specific regulation mapping—comparing pipeline architectures against keyword-based approaches, monolithic LLM prompting, and retrieval-only systems—would clarify the contribution of multi-stage decomposition. Third, the systematic study of LLM-as-judge reliability in legal domains is a prerequisite for credible automated evaluation: our results suggest that legal classification may be inherently harder for LLM judges than open-ended text evaluation, due to the underspecified nature of applicability categories.

References

[1] EUR-Lex. (2025). Statistics on EU legal acts. https://eur-lex.europa.eu

[2] GDPR Enforcement Tracker. https://www.enforcementtracker.com

[3] Regulation (EU) 2024/1689 (AI Act).

[4] Directive (EU) 2022/2555 (NIS2 Directive).

[5] Regulation (EU) 2022/1925 (Digital Markets Act).

[6] Regulation (EU) 2023/2854 (Data Act).

[7] Regulation (EU) 2022/2554 (DORA).

[8] IAPP. (2025). US State Privacy Legislation Tracker.

[9] Personal Information Protection Law (PIPL), China, 2021.

[10] Act on the Protection of Personal Information (APPI), Japan.

[11] Digital Personal Data Protection Act, India, 2023.

[12] Bench-Capon, T. J. M., & Sartor, G. (2003). A model of legal reasoning with cases. *Artificial Intelligence*, 150(1–2), 97–143.

[13] Boella, G., van der Torre, L., & Verhagen, H. (2006). Introduction to normative multiagent systems. *CMOT*, 12(2–3), 71–79.

[14] Governatori, G., et al. (2018). On legal contracts, smart contracts, and blockchain. *AI and Law*, 26(4), 377–409.

[15] Prakken, H., & Sartor, G. (2015). Law and logic. *Artificial Intelligence*, 227, 214–245.

[16] Livermore, M. A., & Rockmore, D. N. (2019). *Law as Data*. Santa Fe Institute Press.

[17] Yao, S., et al. (2023). ReAct: Synergizing reasoning and acting. In *ICLR 2023*.

[18] Lewis, P., et al. (2020). Retrieval-augmented generation. In *NeurIPS 2020*.

[19] Tomasev, N., et al. (2026). Centralized coordination in multi-agent systems. *Preprint, under review*.

[20] Palmirani, M., et al. (2011). LegalRuleML. In *RuleML 2011*.

[21] Hashmi, M., et al. (2018). Business process compliance. *KAIS*, 57(1), 79–133.

[22] Niklaus, J., et al. (2023). LEXTREME. In *Findings of EMNLP 2023*.

[23] Guha, N., et al. (2024). LegalBench. In *NeurIPS 2023 Datasets*.

[24] Nay, J. J., et al. (2024). LLMs as tax attorneys. *AI and Law*, 32, 589–614.

[25] Corazza, M., et al. (2025). Detecting normative change. In *ICAIL 2025*.

[26] Dal Pont, A., et al. (2025). Obligation extraction. In *ICAIL 2025*.

[27] Sapienza, S., & Palmirani, M. (2025). Automated obligation extraction from the AI Act. In *ICAIL 2025*.

[28] Zilli, A., et al. (2024). Normative change detection. In *JURIX 2024*.

[29] Bajwa, G., et al. (2025). Multilingual legal NLP. *Nature HSS Communications*.

[30] Bandara, M., et al. (2025). Engineering failures in LLM agent deployments. *Preprint, under review*.

[31] Stanford HAI. (2025). *AI Index Report 2025*. https://aiindex.stanford.edu/report/

[32] Marino, F., et al. (2026). Computational compliance. *Preprint, under review*.

[33] Videsjorden, A., et al. (2026). Modular agent tasks for AI Act compliance. *Preprint, under review.*

[34] Surani, A., & Ho, D. (2025). STARA. In *ICAIL 2025.*

[35] Jin, H., et al. (2025). Centralized coordination patterns. *Preprint, under review.*

[36] Gui, L., et al. (2025). An error taxonomy for LLM legal reasoning. *Preprint, under review.*

[37] Langfuse. (2025). Why 95% of agent deployments fail. https://langfuse.com/blog/2025-03-agents

[38] Regology. (2025). Global regulatory compliance platform. https://regology.com

[39] Compliance.ai (now Archer). Regulatory change management platform. https://www.archerirm.com

[40] CUBE. (2025). Automated regulatory intelligence. https://www.cube.global