Trustworthy On-Premise AI for Machinery-Regulation Compliance Checking: An E-Government Building Block for Smart Territories

Mario Molinara, Domenico Paolo, Claudio Marrocco, Alessandro Bria, Giuseppe Tomasso
University of Cassino and Southern Lazio, Italy
{m.molinara, domenico.paolo, c.marrocco, a.bria, tomasso}@unicas.it

Giovanni Azzarà
AI-ON S.r.l., Italy
giovanni@ai-on.it

EUt+ European University of Technology, EU

***Abstract*—From January 2027 the EU Machinery Regulation 2023/1230 will apply to all machinery placed on the EU market and, together with the Cyber Resilience Act (CRA), adds essential cybersecurity duties. Checking a machine's technical documentation against this corpus is slow and expertise-intensive, making it a natural candidate for an automated *digital public service* that supports market-surveillance authorities and manufacturers. Such a service, however, processes confidential documents and must therefore be trustworthy—private, GDPR-aware and ideally on-premise—which rules out the naive use of a general-purpose cloud chatbot. We present REGMAC, an explainable retrieval-augmented pipeline that decomposes the regulations into individual items and verifies each of them systematically, yielding a traceable, deterministically scored pre-assessment on open-weight LLMs run locally. On four heterogeneous machines, on-premise open models match the categorical verdict of a frontier cloud model (Claude Opus 4.8) while keeping data local. We frame the tool as an e-government building block for smart territories and relate it to compliance-driven finance.**

***Index Terms*—digital public services, e-government, smart cities, machinery regulation, compliance, large language models, retrieval-augmented generation, on-premise AI, privacy, trustworthy AI.**

I. INTRODUCTION

Regulation (EU) 2023/1230, the Machinery Regulation [1], repeals Directive 2006/42/EC and applies to *all* machinery placed on the EU market from January 2027; jointly with the CRA [2] it introduces essential requirements spanning functional safety and cybersecurity. Demonstrating conformity requires cross-reading a machine's documentation against a large, structured legal corpus—slow, costly and hard to perform consistently. This is not only a manufacturer's problem: for public administrations and market-surveillance bodies, scalable compliance verification is a *digital public service* that underpins a fair single market and, financially, conditions market access, upgrade-investment planning and eligibility for public incentives. In smart territories, dense with connected industrial assets, such a service should be trustworthy by design.

A stakeholder today faces three options: (i) *manual expert review*—accurate but slow and non-scalable; (ii) a *general-purpose cloud chatbot* (Claude, ChatGPT, Gemini) queried ad hoc—powerful, yet unsystematic, non-deterministic and, decisively, incompatible with confidentiality because documents leave the organisation for foreign infrastructure; and (iii) a *structured, retrieval-augmented system* deployable on-premise. The task is hard: a large, heterogeneous corpus, dual safety-and-cyber requirements, diverse manuals, no ground truth, and the need for deterministic, auditable output.

Contributions. (1) We frame machinery-compliance pre-assessment as a *trustworthy e-government service* and make its requirements explicit: privacy, statelessness, data sovereignty, determinism and explainability. (2) We describe REGMAC, an on-premise, explainable RAG pipeline that *systematically decomposes* the Machinery Regulation and the CRA into individual items and verifies each, producing traceable, deterministically scored reports. (3) Through a generator ablation over four open-weight LLMs on four heterogeneous machines, we show that *on-premise open models match a frontier cloud model on the categorical verdict*, and we characterise the resulting privacy/quality/determinism trade-off against the realistic alternatives.

II. RELATED WORK

Smart-city research frames technology-enabled public services along the technology, people and institutions dimensions [5], and trustworthy AI in the public sector is now also a regulatory concern under the AI Act [4]. In the legal domain, RAG and specialised models [6], [7] ground LLMs in authoritative text. LLM output variability is a known issue, partially mitigated by self-consistency [9], motivating our focus on determinism and repeatability. To our knowledge, no prior work offers a privacy-preserving, on-premise, systematically-decomposed compliance pre-assessment for the new Machinery Regulation and the CRA.

TABLE I

REALISTIC ALTERNATIVES AGAINST THE CRITERIA THAT MATTER FOR A PUBLIC-SECTOR SERVICE. ONLY THE ON-PREMISE STRUCTURED PIPELINE IS SIMULTANEOUSLY PRIVATE, SYSTEMATIC AND DETERMINISTIC.

Criterion	Manual expert	Cloud chatbot (Claude/…)	RegMac (on-prem RAG)
Privacy / data sovereignty	High	Low (data abroad)	Configurable, high
Systematic coverage	High, slow	Ad-hoc, gaps	Per-item, complete
Determinism / auditability	Analyst-dep.	Low, black-box	Deterministic, traceable
Domain grounding	Expert	Generalist	2023/1230+CRA+62443
Cost / scalability	High, low scale	Per-token, lock-in	Predictable, scalable

III. METHOD

Problem. Given a machine manual and the regulatory corpus, produce a structured, traceable and reproducible pre-assessment—a triage that flags the main gaps and their priority, not a CE declaration.

RegMac. The regulations are split by legal structure (chapters, articles, annexes, paragraphs) and each item is classified by an LLM (area, applicable roles, machine types, severity, related standards such as IEC 62443 [8]). A machine *profile* is built from the manual; a layered retrieval—mandatory baseline duties, metadata filtering from the profile, and hybrid semantic/lexical search—selects the relevant items, which are merged and de-duplicated. Each retained item is judged for compliance against evidence extracted from the manual, and a *deterministic* rule maps the per-item judgments to a risk band and score; the report cites, for every gap, the legal items and the manual passages that support it. This item-by-item decomposition makes the output systematic, complete and, with the deterministic score, more reproducible than a generic chatbot's free-form answer.

Deployment and privacy. The LLM engine can be placed on-premise (maximum privacy; needs local GPUs), on an Italian/EU cloud (data stay in-country), or on an external cloud with the strongest commercial models (highest quality, lowest privacy); the request is stateless and, on-premise, no document leaves the premises.

Cloud baseline. The strong-but-non-private reference—the closest proxy to "just use a powerful chatbot"—is Claude Opus 4.8 in long-context mode: both regulations and the manual in one request, returning schema-constrained JSON.

IV. EXPERIMENTS AND DISCUSSION

Setup. We use four heterogeneous assets (a variable-frequency drive, a robot controller, an operator HMI, an autonomous mower) and five repetitions per system. For the on-premise pipeline we fix an open-weight embedder ($\sim$8B) and reranker ($\sim$4B) and *ablate the generator* over four open models, reusing the same retrieved context per machine so differences are attributable to the generator alone. Lacking expert ground truth, we report *internal consistency*, not accuracy: within-machine score dispersion, categorical-band agreement and latency. Table I compares the alternatives qualitatively; Table II reports the quantitative summary.

Results. All systems place every machine in the same *yellow* (medium) band (agreement 100%, one model 95%), across both the cloud long-context and the on-premise RAG architectures; since the four assets are all medium-risk, this agreement is encouraging but only weakly informative. The numeric score is model-dependent: RegMac's severity-capped scoring is most conservative ($\approx$38), the LLMs more lenient ($\approx$41–50). One open model anchors near 50 (high repeatability, low discrimination), while another (`qwen3.6-27b`) is repeatable but $\sim$13 min per inference ($\sim$30$\times$ the others), unusable interactively. Overall (Table I), the on-premise pipeline is the only option at once private, systematic and deterministic, while the cloud chatbot maximises raw quality at the cost of privacy and determinism.

TABLE II

SUMMARY OVER FOUR MACHINES AND FIVE REPETITIONS PER SYSTEM. SCORE: MEAN TRAFFIC-LIGHT SCORE (LOWER = MORE SEVERE); STD: MEAN WITHIN-MACHINE DISPERSION (LOWER = MORE REPEATABLE); BAND: MODAL RISK-BAND AGREEMENT. ALL SYSTEMS PLACE EVERY MACHINE IN THE YELLOW BAND.

System	Deploy	Score	Std	Band	Latency
Claude Opus 4.8	cloud	43.1	1.29	100%	55 s
RegMac(Llama-3.3-70B, RAG)	on-prem	38.4	1.62	100%	–
RegMac(`qwen3.5-122b`, RAG)	on-prem	44.9	3.40	100%	48 s
RegMac(`qwen3.6-27b`, RAG)	on-prem	41.5	1.70	100%	13 min
RegMac(`mistral-small-4-119b`, RAG)	on-prem	49.8	0.50	100%	16 s
RegMac(`gpt-oss-120b`, RAG)	on-prem	45.5	2.68	95%	22 s

V. CONCLUSIONS AND FUTURE WORK

On this preliminary benchmark, on-premise open-weight models reproduce the categorical verdict of a frontier cloud model while preserving data locality, statelessness and explainability—the trust properties a public-sector service demands. REGMAC is thus a credible e-government building block for compliance in smart territories, with a direct link to compliance-driven finance (market access, incentive eligibility). The main limitations are the absence of expert ground truth, the small, medium-risk-only benchmark, and the cloud-vs-RAG confound. Future work will add an expert ground truth and finer per-item metrics, a fully systematic report covering every article and annex, cross-engine score calibration, and commercial models to quantify the quality gained by relaxing privacy.

REFERENCES

[1] European Parliament and Council, "Regulation (EU) 2023/1230 on machinery," Off. J. Eur. Union, L 165, 2023.

[2] European Parliament and Council, "Regulation (EU) 2024/2847 (Cyber Resilience Act)," Off. J. Eur. Union, 2024.

[3] European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Off. J. Eur. Union, L 119, 2016.

[4] European Parliament and Council, "Regulation (EU) 2024/1689 (Artificial Intelligence Act)," Off. J. Eur. Union, 2024.

[5] T. Nam and T. A. Pardo, "Conceptualizing smart city with dimensions of technology, people, and institutions," in Proc. 12th Int. Digital Government Research Conf. (dg.o), 2011, pp. 282–291.

[6] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020, pp. 9459–9474.

[7] I. Chalkidis et al., "LEGAL-BERT: The muppets straight out of law school," in Findings of EMNLP, 2020, pp. 2898–2904.

[8] International Electrotechnical Commission, "IEC 62443: Security for industrial automation and control systems," Geneva, 2018–2021.

[9] X. Wang et al., "Self-consistency improves chain-of-thought reasoning in language models," in Proc. ICLR, 2023.