Toward Trustworthy On-Premise Compliance Pre-Assessment of Industrial Machinery: A Preliminary, Proof-of-Concept Comparison of a Cloud LLM and a Custom Open-Weight RAG Pipeline

Mario Molinara[1,2], Domenico Paolo[1,2], Claudio Marrocco[1,2], Alessandro Bria[1,2], Giuseppe Tomasso[1,2], and Giovanni Azzarà[3]

[1] Department of Electrical and Information Engineering, University of Cassino and Southern Lazio, Cassino (FR), Italy
{m.molinara,domenico.paolo,c.marrocco,a.bria,tomasso}@unicas.it
[2] EUt+ European University of Technology, European Union
[3] AI-ON S.r.l., Italy
giovanni@ai-on.it

Abstract. The Machinery Regulation (EU) 2023/1230 will apply to all machinery placed on the EU market from January 2027, introducing, together with the Cyber Resilience Act (EU) 2024/2847, explicit cybersecurity and safety obligations. Manually checking technical documentation against these requirements is costly and error-prone, which motivates automated, document-grounded compliance pre-assessment tools. Such tools, however, must be *trustworthy*: they handle confidential machine documentation and should therefore favour stateless inference (no input retained in the request lifecycle), support GDPR obligations, and ideally be deployed on national or local on-premise infrastructure. This paper is an explicitly *preliminary, proof-of-concept* study comparing two design philosophies for this task: (i) a standard cloud large language model (LLM), Claude Opus 4.8, used in a long-context configuration; and (ii) RegMac, a custom retrieval-augmented (RAG) pipeline running on a local Linux server with open-weight models. Using the same prompt and the executive report structure produced by RegMac, we generate compliance pre-assessments for four heterogeneous industrial assets, with five repetitions each, and we keep the embedding and reranking models fixed while ablating the open-weight *generator* across four candidates. Crucially, we have *no expert ground truth*: we therefore measure *internal consistency* (agreement between systems and run-to-run repeatability), *not* accuracy. On this small benchmark all systems converge on the same categorical risk band—though the benchmark contains no clearly low- or high-risk asset, so this agreement is only weakly informative—while the numeric severity score is model-dependent. We read the study as a system-design and trust-requirements contribution, and outline the substantially larger evaluation (expert labels, finer metrics, more assets, additional commercial models) needed before any claim about which model is most effective.

Keywords: Machinery Regulation (EU) 2023/1230 · Cyber Resilience Act · Compliance pre-assessment · Retrieval-Augmented Generation · On-premise open-weight LLMs · Privacy and unlearning · Explainability · Repeatability.

1 Introduction

Regulation (EU) 2023/1230 (the "Machinery Regulation") [5] repeals the Machinery Directive 2006/42/EC [3] and *applies to all machinery placed on the EU market from January 2027*. Unlike its predecessor, it explicitly addresses digital and connected machinery and, in conjunction with the Cyber Resilience Act (CRA, Regulation (EU) 2024/2847) [6], introduces essential requirements that span both functional/physical safety (control systems, emergency stops, guards and interlocks, moving parts, ergonomics, instructions and residual risks) and *cybersecurity* (protection against corruption of software and safety-related data, secure remote access and maintenance, vulnerability and update management, authentication and access control, logging and traceability, integrity and confidentiality of data). Demonstrating conformity requires cross-reading a machine's technical documentation against a large, structured legal corpus—an activity that is time-consuming, requires scarce expertise, and is difficult to perform consistently.

This setting makes a strong case for *automated compliance pre-assessment*: tools that ingest a machine's manuals, retrieve the relevant normative requirements, and produce a structured, explainable report indicating the main gaps and the priority of further technical investigation. We stress *pre*-assessment: the goal is to triage and prioritise, not to replace a qualified assessor or to issue a CE declaration.

Use-case scenarios. The need for such tools arises in different scenarios, for different reasons: (i) *existing machines* that must be re-evaluated and, where appropriate, *revamped* or substantially modified to meet the new requirements; (ii) *new machines* being placed on the market, which must comply from the outset; and (iii) *machines still in design*, where early detection of gaps is cheapest to act upon (a secure-by-design verification). The same pipeline must therefore adapt the framing of its findings to the work context.

Trust requirements. Because the input is confidential, often proprietary, technical documentation, the tool itself must be trustworthy in the sense targeted by the TrustDOC workshop—privacy, unlearning, robustness and explainability:

- **Statelessness / unlearning.** The inference step should be *stateless*, i.e. retain no input beyond the request lifecycle. We are careful not to overstate this: true "zero-memory" operation is an *architectural goal*, not a property we verify here, and it depends on logging, prompt/KV caching, temporary files and contractual data-retention policies that must all be controlled at deployment. When achieved, statelessness is a strong form of unlearning—there is little to forget because little is kept.

- **Privacy / GDPR.** Sending proprietary manuals to third-party cloud services raises confidentiality and GDPR [4] concerns. On-premise deployment under the data controller's direct control *supports* (but does not by itself *guarantee*) GDPR compliance, which remains an organisational and legal matter beyond the model.
- **Robustness.** For an assessment to be actionable, repeated runs on the same machine should yield consistent verdicts; we quantify this below.
- **Explainability.** Each reported gap must be traceable to specific normative provisions and to the supporting evidence in the manual.

These requirements are in tension with the most convenient option—calling a large, proprietary cloud LLM—which offers state-of-the-art reasoning and very long context windows but limited guarantees on data retention and locality. The alternative is a custom pipeline built on *open-weight* models served on-premise; open models have smaller context windows, so they require retrieval-augmented generation (RAG) [11] to fit the normative corpus, and they are more heterogeneous in their behaviour.

Contribution. This is an explicitly *preliminary, proof-of-concept* study that compares the two philosophies on a common task and a common report structure; it is a system-design and trust-requirements contribution, not a mature experimental evaluation. Its contributions are: (1) a problem framing that ties machinery-compliance pre-assessment to the privacy, unlearning, robustness and explainability themes of trustworthy document understanding; (2) RegMac, a custom on-premise RAG pipeline that produces explainable, chunk-grounded executive pre-assessments, contrasted with a long-context cloud LLM (Claude Opus 4.8); (3) a generator-level ablation over four open-weight LLMs with fixed embedding and reranking models; and (4) a dual qualitative/quantitative evaluation centred on the *repeatability* of the verdict across repetitions. Because we have no expert ground truth and use only four assets, we report internal consistency rather than accuracy; Section 4 makes the limitations central and outlines the substantially larger study required before drawing conclusions about model effectiveness.

2 Methodology

We compare two systems that solve the same task—turning a machine manual into a structured compliance pre-assessment against (EU) 2023/1230 and the CRA—under a common report template, and we then ablate the open-weight generator of the custom system.

2.1 The custom on-premise pipeline (RegMac)

RegMac is a retrieval-augmented system that runs entirely on a local server (see the deployment note in Section 2.4). Its pipeline, sketched in Fig. 1, comprises:

1. **Asynchronous ingestion** of the regulatory PDFs and of the machine manuals.

2. **Legal-aware chunking** that splits the legal texts into structured units (articles, chapters, annexes, paragraphs) rather than fixed windows.
3. **LLM-based chunk classification**, attaching to each normative chunk metadata such as regulatory area, applicable roles, applicable machine types, severity, requirement type and related standards (e.g. IEC 62443 [7], ISO 13849 [9], IEC 62061 [8]), to improve semantic retrieval.
4. **Embedding** of chunks into a vector store and **hybrid retrieval** (dense vector search [10] combined with lexical search and rank fusion [2]), followed by **reranking**.
5. **Machine profiling and applicability classification**: a profile of the machine (digital elements, remote access, wireless/cloud connectivity, AI, safety control system, moving parts, lifting, ...) is built from the manuals; each candidate requirement is then classified as applicable, uncertain or not applicable to that profile.
6. **Requirement grouping, evidence retrieval and batch judgment**: the applicable requirements are grouped and, for each group, the relevant manual evidence is retrieved; the generator LLM produces, for each requirement, a judgment in {COMPLIANT, PARTIAL, NON_COMPLIANT, NOT_ASSESSABLE} with the cited evidence.
7. **Deterministic scoring and explainable report**: a compliance score (0–100) and a qualitative risk level are computed by a fixed, deterministic rule from the judgments; the executive report renders a traffic-light synthesis, the main operational gaps (each with *operational risk*, *potential impact*, *recommended action* and a normative reference) and a mapping from gaps to mitigation building blocks, plus an appendix listing the analysed documents.

The traffic-light percentage shown in the report is $w = \min\big(s,\, c(\sigma)\big)$, where s is the compliance score and $c(\sigma)$ is a cap determined by the most severe gap ($c = 25$ for CRITICAL, 50 for HIGH, 66 for MEDIUM, 80 otherwise); the colour is red for a CRITICAL risk level, yellow for HIGH/MEDIUM and green otherwise. Because scoring is deterministic given the judgments, residual numeric variability stems only from the LLM judgments.

2.2 The cloud baseline (Claude Opus 4.8)

As a representative state-of-the-art cloud system we use Claude Opus 4.8. Thanks to its very large context window it is used in a *long-context* configuration that does *not* require retrieval: the full text of both regulations and the machine manual are provided in a single request, together with the same role and instructions used by RegMac, and the model is asked to return the executive report as a structured JSON object. We enable adaptive reasoning, request structured (schema-constrained) output, and rely on prompt caching of the (identical) regulatory corpus across requests. Calls are stateless at the API level. This configuration represents the "just use a strong cloud model" option against which a custom pipeline must justify itself.

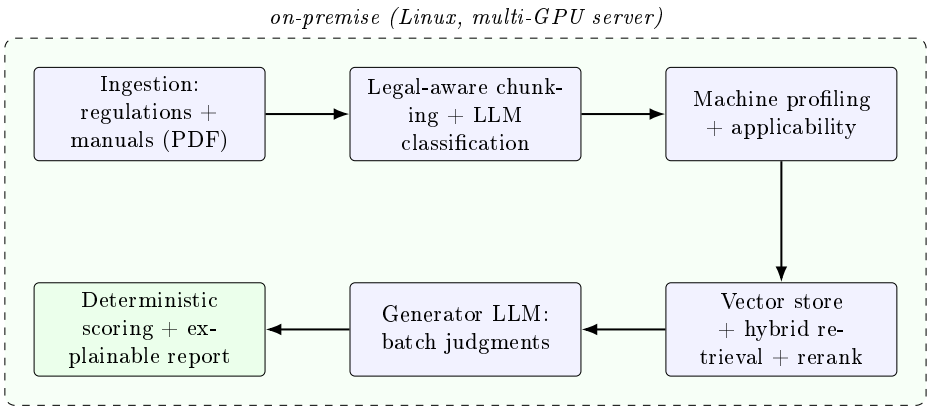

Fig. 1. The custom on-premise RegMac pipeline. The dashed region runs entirely on a local Linux server; by design, documents stay on the controller's infrastructure and the pipeline aims at stateless operation (subject to the deployment controls discussed in the text).

2.3 Generator ablation for the open-weight system

A key methodological point is *where* model choice matters. In a RAG pipeline the *embedding* model (used to retrieve the relevant normative chunks) and the *reranker* (used to reorder candidates) jointly determine *what evidence the generator sees*; they are therefore important but are not the component that writes the verdict. To isolate the effect of the *generator*, we keep a single open-weight embedding model (`Qwen3-Embedding-8B`, 4096-dimensional embeddings) and a single open-weight reranker (`Qwen3-Reranker-4B`) *fixed*, and we ablate the generator across four open-weight LLMs of different families and sizes (Section 3.1). Crucially, for a given machine the retrieved normative context is computed *once* and reused for all generators, so that any difference in the output is attributable to the generator alone. Fig. 2 summarises the experimental comparison.

2.4 On-premise infrastructure

The custom system and all open-weight models run on a *local* server: a Linux-based machine built on standard (x86-64) server hardware and equipped with multiple data-center GPUs providing sufficient aggregate VRAM (on the order of a few hundred GB) to serve open-weight LLMs up to about 122B parameters, together with the embedding and reranking models. This makes it possible to keep the entire workflow—ingestion, retrieval and generation—inside the data controller's premises. Such a deployment *supports* GDPR obligations and a stateless operating mode, but does not by itself guarantee them: as noted in Section 1, logging, prompt/KV caching, temporary files and contractual retention must still be controlled, and verifying these properties is out of scope here.

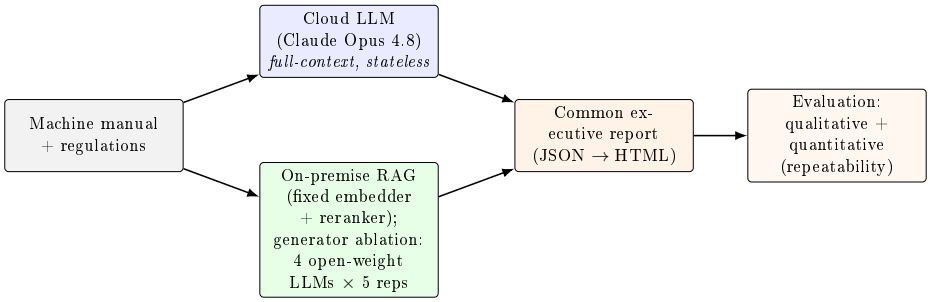

Fig. 2. Experimental comparison. Both philosophies produce the same executive report structure; the on-premise system additionally undergoes a generator-level ablation with a fixed retrieval front-end. Outputs are compared qualitatively and quantitatively, with emphasis on repeatability across the five repetitions per machine.

2.5 Evaluation modalities

We evaluate the systems both qualitatively and quantitatively. Each system produces five independent repetitions per machine; the cloud model is driven with adaptive reasoning, while the open-weight generators use a moderate sampling temperature, so that the five repetitions probe the natural variability of each system, mirroring self-consistency analyses of LLM sampling [16].

Quantitative evaluation (repeatability/determinism). The central quantitative question is how *deterministic*, or at least *reasonably repeatable*, a system is on a single machine. For each (system, machine) pair we measure: (a) the within-machine standard deviation and range of the traffic-light score w across the five repetitions; (b) the *risk-band agreement*, i.e. the fraction of repetitions assigned the modal colour band; and (c) the variability of the number of reported gaps. We also record indicative inference latency. Lower score dispersion and higher band agreement indicate a more repeatable, and thus more dependable, system.

Qualitative evaluation. We assess each system along six axes that we consider decisive for trustworthy, actionable pre-assessment: (1) *normative grounding*—whether gaps cite specific provisions of (EU) 2023/1230 and the CRA (annexes, points) rather than generic statements; (2) *faithfulness to the manual*—whether claims are tied to evidence actually present in the documentation, with low hallucination; (3) *safety+cyber coverage*—whether both physical/functional safety and cybersecurity are addressed; (4) *output-schema conformance*—whether the model reliably emits the required structured format; (5) *explainability/traceability*—whether findings can be traced back to sources; and (6) *output style*—conciseness vs verbosity and the resulting executive readability.

Table 1. Indicative *native* context windows of the language models (from the vendors' documentation). Claude's 1M window is the one offered on the Anthropic API used here. For the open models we used a single, smaller *effective* window ($\leq$128K) shared across the ablation, both for practical serving reasons and to keep the comparison fair; in all cases the full $\approx$190K-token regulatory corpus, plus a manual, makes retrieval the pragmatic choice, whereas Claude's very large window allows full-context use.

Model	Deployment	Context (native, indic.)
Claude Opus 4.8	cloud	$\sim$1,000K (API)
`Llama-3.3-70B`	on-premise	$\sim$128K
`qwen3.5-122b`	on-premise	$\sim$256K
`qwen3.6-27b`	on-premise	$\sim$256K
`mistral-small-4-119b`	on-premise	$\sim$256K
`gpt-oss-120b`	on-premise	$\sim$128K

3 Experiments

3.1 Setup

Assets. We use four heterogeneous industrial assets, each documented by its manufacturer's manual: an ABB ACS880-01 variable-frequency drive, an ABB OmniCore C30 robot controller, a Pro-face GP-4100 series operator HMI, and a Segway Navimow H autonomous robotic mower. They differ in connectivity, presence of safety-related control systems and digital/cyber exposure, which stresses the applicability logic of the pipeline.

Models. The cloud baseline is Claude Opus 4.8 [1] in the long-context configuration of Section 2.2. The on-premise system uses a fixed embedder (`Qwen3-Embedding-8B`) and a fixed reranker (`Qwen3-Reranker-4B`) [15], and ablates the generator over four open-weight LLMs spanning families and sizes: a 122B model (`qwen3.5-122b`) [15], a 27B reasoning model (`qwen3.6-27b`) [15], a $\sim$119B Mistral model (`mistral-small-4-119b`) [13], and a 120B open reasoning model (`gpt-oss-120b`) [14]. The figures produced by the original RegMac system (whose generator is the open 70B `Llama-3.3-70B` [12] instruction model) are included as a reference. Table 1 reports the indicative native context windows: Claude's very large window makes full-context use possible, whereas for the open models we adopt a single, smaller effective window across the ablation, so that retrieving the relevant chunks—rather than feeding the whole $\approx$190K-token corpus plus a manual—is the pragmatic and fair choice.

Prompting and output handling. All systems receive the same expert role, the same processing instructions and a common target schema. For the open-weight generators, the relevant normative context (top-k retrieved chunks) and the manual are placed in the prompt, and the model is asked to fill an explicit JSON skeleton. Because open models do not support constrained decoding [17] as reliably as the cloud API, their outputs are passed through a tolerant normaliser

Table 2. Mean traffic-light score (%) over five repetitions per machine, with [min–max] range. Lower scores denote a more conservative (more severe) assessment. All entries are in the yellow band; the single red outlier produced by `gpt-oss-120b` on the HMI is discussed in the text.

System	ACS880 (drive)	OmniCore (robot)	GP-4100 (HMI)	Navimow (mower)
Claude Opus 4.8 (full-context)	42.0 [42–42]	49.6 [48–50]	36.0 [34–38]	45.0 [42–48]
RegMac(Llama-3.3-70B, RAG)	38.2 [36–40]	41.6 [37–44]	36.0 [34–37]	37.8 [36–40]
RegMac(`qwen3.5-122b`, RAG)	42.0 [35–45]	45.6 [35–50]	42.0 [35–45]	50.0 [50–50]
RegMac(`qwen3.6-27b`, RAG)	41.0 [35–45]	45.0 [45–45]	35.0 [35–35]	45.0 [42–48]
RegMac(`mistral-small-4-119b`, RAG)	50.0 [50–50]	50.0 [50–50]	50.0 [50–50]	49.0 [45–50]
RegMac(`gpt-oss-120b`, RAG)	45.0 [45–45]	48.0 [45–50]	41.0 [30–45]	48.0 [45–50]

that maps heterogeneous key names (including Italian variants and nested objects) onto the canonical schema; the resulting JSON is rendered into the same HTML report as the cloud and RegMac outputs. We generate five repetitions per machine and per system.

Reproducibility settings. For transparency we record the main settings. Retrieval uses character-based chunks of ∼1500 characters with 200 overlap and returns the top $k = 24$ chunks per machine; the manual is truncated to ∼80K characters. The open-weight generators are sampled at temperature 0.6 with *no fixed random seed* (so the five repetitions reflect genuine sampling variability) and a generation budget of ∼14K tokens; Claude is driven with adaptive reasoning and schema-constrained JSON output. Claude is the `claude-opus-4-8` model on the Anthropic API; the open models are served through an OpenAI-compatible local endpoint, and their free-form JSON is coerced to the canonical schema by a tolerant normaliser. We deliberately do *not* claim full reproducibility: the exact quantization, serving framework and per-call API versions are part of the deployment and are not pinned here; reporting them, together with seeds, in a reproducibility appendix is part of future work.

3.2 Quantitative results

Table 2 reports, for every system and machine, the mean traffic-light score over five repetitions with its observed range. All systems and all machines fall in the yellow (medium) band, with a single exception (one red outlier discussed below). Table 3 summarises repeatability and indicative efficiency.

Three observations stand out. First, the *categorical verdict is highly repeatable and consistent across very different systems*: every system places every machine in the yellow band, and band agreement is 100% everywhere except `gpt-oss-120b` (95%, due to a single red outlier on the HMI). This is encouraging for automation: the coarse risk classification appears robust to the choice of model and even to the choice of architecture (long-context vs RAG). We caution, however, that this agreement is only *weakly informative*: our four assets are all medium-risk, so the benchmark contains no clearly low- or high-risk case that would force a green or red verdict; band stability is therefore close

Table 3. Repeatability and indicative efficiency. "Score std" and "range" are averaged over machines (within-machine, across the five repetitions); "band agreement" is the fraction of repetitions assigned the modal colour band; latency is per inference. Lower std and higher agreement mean more repeatable.

System	Score std	Range	Band agr.	Mean gaps	Latency
Claude Opus 4.8	1.29	3.0	100%	5.0	∼55 s
RegMac(Llama-3.3-70B, RAG)	1.62	4.5	100%	3.0	–
RegMac(`qwen3.5-122b`, RAG)	3.40	8.8	100%	3.9	∼48 s
RegMac(`qwen3.6-27b`, RAG)	1.70	4.0	100%	4.0	∼13 min
RegMac(`mistral-small-4-119b`, RAG)	0.50	1.2	100%	5.0	∼16 s
RegMac(`gpt-oss-120b`, RAG)	2.68	6.2	95%	4.8	∼22 s

to expected, and a discriminative benchmark with calibrated ground truth is needed to read it as evidence of correctness. Second, the *numeric score is model-dependent.* RegMac is the most conservative (mean ≈ 38) and the most concise (three executive gaps), reflecting its deterministic, severity-capped scoring; the cloud and open-weight LLMs tend to be somewhat more lenient (means ≈ 41–50) and more granular (four to five gaps). Third, repeatability differs sharply between generators: `mistral-small-4-119b` shows almost no dispersion (std 0.50) but only because it *anchors the score to* 50 *regardless of the machine*, i.e. its stability reflects low discriminative power rather than genuine determinism; Claude is both stable and informative (std 1.29 with machine-dependent scores), whereas `qwen3.5-122b` is the most variable (std 3.40, range up to 8.8). Finally, `qwen3.6-27b`, a reasoning model, completed all four assets but was impractically slow on the local server (∼13 min per inference, i.e. roughly 30× the other open models); despite its good repeatability, this latency makes it unsuitable for an interactive, on-premise tool—itself a relevant finding.

3.3 Qualitative results

Table 4 summarises the qualitative comparison. Both Claude and RegMac provide strong normative grounding: across machines they cite specific provisions (e.g. Annex III points of (EU) 2023/1230 and Annex I/II points of the CRA) and tie gaps to concrete manual evidence (for the drive, for instance, firmware-update and removable-memory handling, the wireless control panel, and fieldbus interfaces such as PROFINET/Modbus). RegMac additionally grounds each finding in a controlled set of retrieved legal chunks and renders a fixed executive structure, which is the strongest position on explainability and schema conformance. The open-weight generators are competent on grounding and coverage but heterogeneous on output discipline: schema conformance was the main practical obstacle, as several models emitted extra or renamed keys, nested the score fields, or were truncated, and therefore required a normalisation step that the constrained-decoding cloud API made unnecessary. `mistral-small-4-119b` is fast and well-structured but its near-constant score limits its usefulness as a discriminator; `gpt-oss-120b` offers the best quality/latency trade-off among the

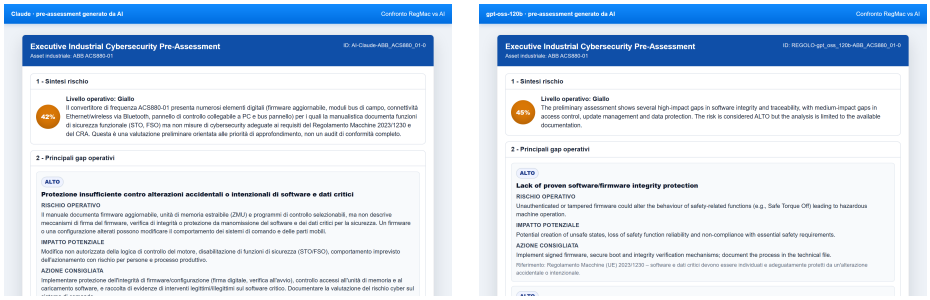

Fig. 3. Top of the executive pre-assessment for the ABB ACS880-01 drive: Claude Opus 4.8 (left) and RegMac(`gpt-oss-120b`, RAG) (right). Both follow the same structure—risk synthesis (traffic light) followed by the main operational gaps, each with operational risk, potential impact, recommended action and a normative reference. Claude answers in Italian, as instructed, and is more concise; the open model is denser and, here, replied in English.

open models but showed the only cross-band instability and, notably, answered in English despite the Italian instruction. Fig. 3 shows the top of the report for the ABB ACS880-01 drive produced by Claude and by the most verbose open model: the two share the same executive structure, while differing in language, conciseness and output discipline.

To make the grounding concrete, one gap from the Claude report on the ABB ACS880-01 drive reads, in essence:

- **Requirement.** Reg. (EU) 2023/1230, Annex III §1.2.1 (safety and reliability of control systems); CRA Annex I, Part I §2(d)–(e); IEC 62443-3-3.
- **Manual evidence.** The ACS-AP-W control panel (Bluetooth), the PC connection via Drive Composer, the panel bus and the fieldbus modules (PROFINET, EtherNet/IP, Modbus TCP) are documented as interfaces, but the manual describes no authentication, access control or encryption for them.
- **Gap / action.** Remote access and connectivity are not governed on the cyber side; define authentication and access control, disable unnecessary wireless interfaces, segment the OT network and encrypt teleservice channels.

This requirement–evidence–action triple is exactly what a reviewer can audit, and is the unit on which the finer agreement metrics proposed in Section 4 would operate.

4 Discussion, Conclusions and Future Work

Findings (read with care). On this small benchmark the coarse compliance verdict (the colour band) is stable across architectures and models, which is a *hint*—not proof—that automated pre-assessment could produce a dependable triage

Table 4. Qualitative comparison. H/M/L denote high/medium/low. Grounding: specificity of normative references; Faithfulness: adherence to manual evidence; Cov.: joint safety+cyber coverage; Schema: reliability of the structured output without post-processing; Explain.: traceability of findings.

System	Ground.	Faith.	Cov.	Schema	Explain.	Style
Claude Opus 4.8	H	H	H	H	M–H	detailed (5 gaps)
RegMac(Llama-3.3-70B, RAG)	H	H	H	H	H	concise, executive (3 gaps)
RegMac(`qwen3.5-122b`, RAG)	H	H	M–H	L	M	variable, IT keys
RegMac(`qwen3.6-27b`, RAG)	M–H	H	M	M	M	repeatable but very slow
RegMac(`mistral-small-4-119b`, RAG)	M	M	H	L	M	verbose; score anchored to 50
RegMac(`gpt-oss-120b`, RAG)	M–H	M–H	H	M–H	M	fast; replied in English; one red outlier

signal even with on-premise open-weight models. As stressed in Section 3.2, the four assets are all medium-risk, so this stability is close to expected and must not be read as accuracy. The finer numeric severity is model-dependent and not comparable across systems without calibration; the deterministic scoring of RegMac is an advantage here, as it makes the numeric output reproducible by construction given the model judgments. The cloud model is strong and convenient (long context removes the need for retrieval, constrained decoding guarantees the schema) but offers weaker guarantees on data locality and retention; the on-premise pipeline can match the categorical verdict and better supports the trust properties the application demands—stateless inference, local deployment and explainable, chunk-grounded findings—at the cost of a retrieval front-end and per-model output-discipline issues.

Relation to trustworthy document understanding. The *design* maps onto the TrustDOC themes, while we are careful about what is actually demonstrated. *Privacy* is supported by keeping proprietary manuals on local infrastructure; *unlearning* is targeted through stateless inference (an architectural goal, subject to the deployment controls of Section 1, not a verified property); *robustness* is what our repeatability analysis measures; and *explainability* is built into the chunk-grounded, structured report.

Limitations (central to this work). This is a preliminary, proof-of-concept study and its conclusions are correspondingly limited. (i) **No ground truth**: without expert compliance labels we measure internal consistency and inter-system agreement, *not* accuracy; the numbers say how *repeatable* and *mutually consistent* the systems are, not how *correct*. (ii) **Weak discriminative power**: the four assets are all medium-risk, so the 100% band agreement is nearly trivial and the benchmark cannot reveal green/red disagreements. (iii) **Unfair by construction**: the cloud model is used in full-context mode while the open models use RAG, so the comparison conflates the model with its information-access strategy, and the severity scales are not calibrated across systems. (iv) **Small scale**: four machines, five repetitions. (v) **Partial reproducibility**: quantization, serving

framework, API versions and seeds are not pinned. (vi) **Instruction following**: one open model answered in English despite the Italian instruction. These limitations, not the headline numbers, are the main message of the paper.

Future work. We plan to (i) build an expert-annotated ground truth and report accuracy, calibration and inter-annotator agreement; (ii) add *finer* agreement metrics beyond the colour band—per-gap agreement, overlap of the cited normative provisions, similarity of the recommended actions, and a manual expert rating on a subset—operating on the requirement–evidence–action triple illustrated above; (iii) enlarge and *diversify* the benchmark so that it contains clearly low- and high-risk assets, and analyse the three use-case scenarios (new, revamping, in-design) separately; (iv) ablate the embedding and reranking models, not only the generator, and test the cloud model under RAG and the open models under longer contexts to disentangle model from access strategy; (v) report a full reproducibility appendix (quantization, serving framework, API versions, seeds); and (vi) add further *commercial* models from other vendors (e.g. OpenAI and Google) and a thorough evaluation before any claim about which system is most effective. These extensions are what would turn this preliminary proof of concept into a comprehensive, reproducible benchmark for trustworthy, on-premise machinery-compliance pre-assessment.

References

1. Anthropic: Claude opus 4.8 — model overview and documentation. https://docs.anthropic.com/en/docs/about-claude/models (2026), context window 1M tokens on the Anthropic API. Last accessed 2026
2. Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of ACM SIGIR. pp. 758–759 (2009), dOI:10.1145/1571941.1572114
3. European Parliament and Council: Directive 2006/42/EC of the european parliament and of the council on machinery (Machinery Directive). Official Journal of the European Union, L 157 (2006), https://eur-lex.europa.eu/eli/dir/2006/42/oj
4. European Parliament and Council: Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L 119 (2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj
5. European Parliament and Council: Regulation (EU) 2023/1230 of the european parliament and of the council of 14 june 2023 on machinery. Official Journal of the European Union, L 165 (2023), https://eur-lex.europa.eu/eli/reg/2023/1230/oj
6. European Parliament and Council: Regulation (EU) 2024/2847 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). Official Journal of the European Union, L 2024/2847 (2024), https://eur-lex.europa.eu/eli/reg/2024/2847/oj
7. International Electrotechnical Commission: IEC 62443-3-3, 62443-4-1 and 62443-4-2 — security for industrial automation and control systems: system security requirements and security levels, secure product development life-cycle, and technical security requirements for components. IEC, Geneva (2013–2019)
8. International Electrotechnical Commission: IEC 62061:2021 — safety of machinery: Functional safety of safety-related control systems. IEC, Geneva (2021)

9. International Organization for Standardization: ISO 13849-1:2023 — safety of machinery: Safety-related parts of control systems — part 1: General principles for design. ISO, Geneva (2023)
10. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of EMNLP. pp. 6769–6781 (2020), arXiv:2004.04906
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 9459–9474 (2020), arXiv:2005.11401
12. Meta AI: Llama-3.3-70B-Instruct model card. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct (2024), last accessed 2026
13. Mistral AI: Mistral small model card. https://huggingface.co/mistralai (2025), native context window up to 256K tokens. Last accessed 2026
14. OpenAI: gpt-oss-120b open-weight model card. https://huggingface.co/openai/gpt-oss-120b (2025), runs on a single 80 GB GPU. Last accessed 2026
15. Qwen Team: Qwen3 model family (embedding, reranker and instruction models). https://huggingface.co/Qwen (2025), last accessed 2026
16. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain-of-thought reasoning in language models. In: International Conference on Learning Representations (ICLR) (2023), arXiv:2203.11171
17. Willard, B.T., Louf, R.: Efficient guided generation for large language models. arXiv preprint arXiv:2307.09702 (2023)