

## Article

# Traceable Cybersecurity Conformity Pre-Assessment of Electric Vehicle Charging Equipment Using Cross-Standard Gap Analysis and a Local LLM

Jin-hyeok Kang <sup>1,2,\*</sup> , Seung-hwan Lee <sup>1</sup> and You-Suk Bae <sup>2</sup><sup>1</sup> Korea Testing Laboratory, Seoul 08389, Republic of Korea; leesh@ktl.re.kr<sup>2</sup> Department of Smart Manufacturing Engineering, Tech University of Korea, Siheung 15073, Republic of Korea; ysbae@tukorea.ac.kr

\* Correspondence: kang88@ktl.re.kr

## Abstract

Electric vehicle charging equipment now links vehicles, payment and roaming services, cloud platforms, and the grid, but its cybersecurity requirements are fragmented across standards. This study combines a cross-standard gap analysis with a traceable conformity pre-assessment framework for a local large language model; the framework is a design-validation proof of concept, not a production-ready conformity-assessment system. Using EN 18031 as the reference axis, IEC 62443-4-2, ETSI EN 303 645, and NIST IR 8473 requirements were mapped on a five-level correspondence scale by a single researcher and consolidated into integrated requirements with applicability, evidence, and acceptance rules. A clause-aware retrieval-augmented generation pipeline (semantic retrieval only) was compared with no-retrieval and fixed-length baselines on 80 synthetic design-gold cases, reported on the 44 holdout cases not used for tuning. Better retrieval did not ensure trustworthy conformity reasoning: retrieved clause text in the judgment prompt lowered verdict accuracy (0.89 without, 0.77 with), the most accurate configuration (0.93) used retrieval only for citations, and a rule-based classifier with no model or corpus reached 0.66 on the same cases. Citation extraction, citation verification, and human review were each necessary. Extraction guarantees citation existence, but exact-gold-clause agreement remained at most 0.38. External validity on real manufacturer documents is untested; the framework complements rather than replaces accredited certification.

**Keywords:** electric vehicle supply equipment; cybersecurity standards; standards gap analysis; conformity pre-assessment; local large language model; retrieval-augmented generation; citation verification; traceability



Academic Editor: Michael Fowler

Received: 17 August 2026

Revised: 16 September 2026

Accepted: 22 September 2026

Published: 29 September 2026

**Copyright:** © 2026 by the authors.

Published by MDPI on behalf of the World Electric Vehicle Association.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

The global electric vehicle (EV) transition is being accompanied by the rapid deployment of public and private charging infrastructure. More than 1.3 million public charging points were added worldwide in 2024, bringing the installed total above five million [1], and deployment and optimization of the charging infrastructure continue to accelerate worldwide [2]. This expansion changes charging equipment from an isolated power appliance into a distributed cyber-physical endpoint connected to vehicles, charging-station management systems, payment and roaming platforms, cloud services, and the electric grid. A compromise can therefore affect not only personal information and billing but also charging availability, vehicle operation, and distribution-system stability [3,4].

The European regulatory context makes this issue immediately relevant to product conformity. Directive 2014/53/EU established essential requirements for radio equipment [5], and Delegated Regulation (EU) 2022/30 activated the network-protection, privacy, and fraud-prevention requirements of Article 3(3)(d)–(f) for specified radio equipment [6]. The application date was deferred to 1 August 2025 by Delegated Regulation (EU) 2023/2444 [7]. The EN 18031 series was subsequently developed as an assessment-oriented harmonized route for demonstrating these requirements [8–10], subject to the limitations published in Implementing Decision (EU) 2025/138 [11]. The wider Cyber Resilience Act (CRA) now creates a transition to a horizontal product-cybersecurity regime [12]; Delegated Regulation (EU) 2026/339 repeals Delegated Regulation (EU) 2022/30 with effect from 11 December 2027 [13]. Manufacturers consequently need to retain both clause-level evidence and the regulatory status of the standard version used for an assessment.

No single cybersecurity document describes the complete EV-charging problem at the same level of abstraction. EN 18031 addresses internet-connected radio equipment and its conformity-assessment evidence. IEC 62443-4-2 specifies technical security requirements for industrial automation and control components [14]. ETSI EN 303 645 provides a consumer Internet-of-Things baseline [15]. NIST IR 8473 applies the NIST Cybersecurity Framework to the extreme-fast-charging ecosystem [16]. Their scopes, actors, normative force, requirement granularity, and verification approaches differ. A manufacturer can therefore know that several standards are relevant while still lacking a reproducible method for deciding which product evidence satisfies which requirement, where coverage is only partial, and which EV-specific risks remain outside the standards.

Existing EV-charging cybersecurity studies have made substantial progress on protocol threats, certificates, transport security, penetration testing, and power-system consequences. These studies are essential, but they do not by themselves convert multiple standards into an auditable product-evidence workflow. Likewise, a cross-reference table that marks two clauses as related does not indicate whether their strength is equivalent, whether the same asset and actor are covered, or whether the relationship can be verified from product documents and tests. The practical gap lies between standards comparison and conformity preparation.

Large language models (LLMs) appear useful for interpreting standards and technical evidence, but conformity work is an unusually strict setting. A fluent verdict is not useful if it cites a nonexistent clause, silently uses the wrong standard version, mistakes missing evidence for nonconformity, or sends confidential product documents to an external service. Retrieval-augmented generation (RAG) can provide source context, yet retrieval quality, verdict correctness, and citation validity are distinct properties that must be measured separately.

We therefore developed and evaluated a traceable, locally executed LLM framework for the cybersecurity conformity pre-assessment of EV charging equipment. Traceability is used in a verifiable sense. Every automated statement carries identifiers (standard, version, clause locator, and product-evidence locator) that a program or a reviewer can check against the corpus; it does not imply that the cited clause is the most relevant one or that it supports the verdict, which remains a human review task (Section 4.5). Its contributions are as follows:

- A five-level correspondence analysis maps 44 unique EN 18031 requirement identifiers to IEC 62443-4-2, ETSI EN 303 645, and NIST IR 8473, exposing 132 auditable cross-standard relationships rather than binary matches;
- Eight EV-charging threat scenarios connect standard-level gaps to device, operational, and ecosystem requirements, product evidence, assessment methods, and acceptance rules;

- A clause-aware local-RAG architecture preserves hierarchy, normativity, version, cross-references, citation locators, and regulatory status while constraining verdicts to PASS, PARTIAL, FAIL, N/A, or INSUFFICIENT EVIDENCE. Only clause-aware segmentation with semantic retrieval was evaluated; hybrid identifier/keyword retrieval and reranking remain design specifications;
- An evaluation on 80 balanced synthetic judgment cases with deterministic design-gold labels, reported on the 44 holdout cases not used for tuning, and on 24 retrieval queries separates retrieval, verdict, citation, safety, and stability outcomes, with diagnostic follow-up experiments on execution path, a rule-based classifier, extractive citation, verdict and citation path separation, a second model family, citation verification, and repeatability;
- The experiments identify a design separation: verdict accuracy was highest when no retrieved clause text entered the judgment prompt, whereas citation existence was guaranteed only when citations were extracted programmatically from retrieved clause cards rather than written by the model. Even then, exact-gold-clause agreement was at most 0.38, and extraction attached no citation to a third of the definitive verdicts of the most accurate condition. So, the study establishes design requirements and failure modes rather than production readiness.

The system supports preparation and evidence-gap analysis; it does not issue regulatory decisions, replace a notified body or accredited conformity-assessment organization, or convert research-proposed gap controls into legally binding requirements. Final scope, evidence sufficiency, risk acceptance, and conformity decisions remain human responsibilities. Two limitations frame every result: the mapping was performed by one researcher without independent adjudication, and the gold labels of the synthetic cases were derived from the case design rather than from independent expert judgment (Sections 3.2, 3.6 and 5.4).

## 2. Regulatory and Research Background

### 2.1. EV Charging Infrastructure as a Cyber-Physical System

EV charging infrastructure comprises the EV, electric vehicle supply equipment (EVSE), the charging station management system (CSMS), cloud and roaming services, payment systems, and utility or distribution system operator (DSO) interfaces. Vehicle-to-charger communication is governed by ISO 15118 and IEC 61851 families [17,18], while EVSE-to-CSMS communication commonly uses the Open Charge Point Protocol (OCPP) [19]. IEC 62351 provides complementary security mechanisms for power-system data and communications [20]. These link cross-field, operational, and external-grid trust boundaries, as summarized in Figure 1.

The attack surface is correspondingly layered. Network-security surveys of cyber-physical systems classify attacks by attack space, location, and stealthiness and extend the treatment from threat identification to resilient design [21]. A Plug and Charge implementation depends on certificate and Transport Layer Security (TLS) behavior [22]. OCPP research has identified authentication, authorization, replay, denial-of-service, firmware, and backend threats [23–26], and charging-management security frameworks for the wider EV ecosystem have been proposed in this journal [27]. At the device level, however, protocol protection cannot ensure secure boot, confidential-key storage, local-interface control, safe recovery, or physical tamper evidence. At the ecosystem level, it also cannot allocate responsibility for coordinated load attacks, vehicle-to-grid (V2G) settlement manipulation, supply-chain incidents, or grid-synchronized recovery; recent V2G reviews similarly identify governance and standards coordination as open challenges [28]. The relevant unit of analysis is therefore broader than one protocol and narrower than an enterprise-only

governance framework. It is a traceable chain from a normative clause and identified gap to a scoped product or operational evidence item.

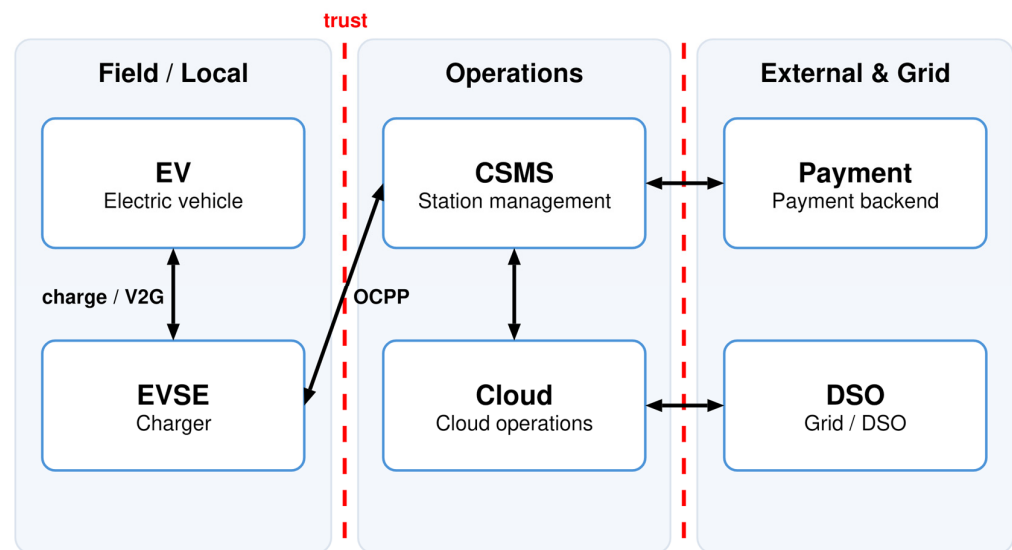


**Figure 1.** Charging-infrastructure system model and trust boundaries.

## 2.2. Standards Selected for Cross-Standard Analysis

The standards were selected because one charger can simultaneously be radio equipment, an industrial control component, a consumer-facing Internet of Things (IoT) device, and part of critical energy infrastructure. Table 1 summarizes the distinct analytical role of each source. EN 18031 was selected as the reference axis because it is directly linked to the Radio Equipment Directive (RED) essential requirements and includes informative mappings to IEC 62443-4-2 and ETSI EN 303 645. NIST IR 8473 was added because it is the only selected source explicitly organized around the EV extreme-fast-charging ecosystem.

**Table 1.** Standards corpus and role in the analysis.

| Source                 | Primary Scope                                                    | Role in This Study                                                                                     |
|------------------------|------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| EN 18031-1/-2/-3:2024  | Internet-connected radio equipment; data/privacy; monetary value | Reference axis; assessment-oriented requirements linked to RED Article 3(3)(d)–(f)                     |
| IEC 62443-4-2:2019     | Technical security requirements for IACS components              | Device hardening, availability, authentication, integrity, and component-level controls                |
| ETSI EN 303 645 V3.1.3 | Consumer IoT baseline security                                   | Consumer protection, vulnerability disclosure, support period, update, privacy, and usability controls |
| NIST IR 8473:2023      | EV extreme-fast-charging ecosystem profile                       | Cross-domain governance, supply chain, incident response, recovery, utility, and partner coordination  |

EN 18031 organizes requirements by mechanism, such as access control (ACM), authentication (AUM), software update (SUM), secure storage (SSM), secure communication (SCM), confidential cryptographic keys (CCK), cryptography (CRY), general equipment capabilities (GEC), resilience (RLM), network monitoring (NMM), traffic control (TCM), logging (LGM), deletion (DLM), and user notification (UNM). The standard first determines applicability and then assesses the appropriateness of an implementation through documented information, conceptual assessment, functional-completeness assessment, and functional-sufficiency assessment [8–10]. This assessment orientation makes EN 18031 useful as a traceability backbone, but it does not make adjacent standards substitutable. Its Annex B expressly notes that compliance with IEC 62443-4-2 does not itself establish EN 18031 conformity.

IEC 62443-4-2 derives component requirements from seven foundational requirements defined in the IEC 62443 series [29]. ETSI EN 303 645 combines mandatory and recommended consumer-IoT provisions; its associated technical specification and assessment methodology provide implementation and testing support [30,31]. NIST IR 8473 applies the five functions of NIST Cybersecurity Framework (CSF) 1.1 to EV, EVSE, cloud/third-party, and utility/building domains [32]. The NIST profile is therefore broad but often more abstract than a device-level test requirement. Prior work has mapped industrial IoT technologies to IEC 62443 [33], but a binary mapping does not distinguish a direct, testable equivalent from a high-level thematic relationship.

### 2.3. LLMs, RAG, and Traceable Conformity Reasoning

Transformer-based language models [34–37] can interpret natural-language requirements and heterogeneous product evidence, while RAG supplies retrieved source passages to knowledge-intensive generation [38,39]. Dense retrieval and sentence embeddings support semantic matching between differently worded queries and passages [40,41]. This approach is well-suited to standards analysis, where the evidence consists largely of text, tables, definitions, applicability conditions, and documentary records.

General-purpose RAG is nevertheless insufficient for conformity reasoning. A fixed-length chunk may separate a requirement from its scope, decision tree, normative verb, table, annex, or cross-reference. A retrieved paragraph may be relevant but belong to the wrong version. A generated citation may look plausible while having no exact counterpart in the corpus. In addition, product design files and test reports may be commercially sensitive. The framework therefore uses local execution and clause-aware retrieval, and it evaluates citation validity independently of answer correctness.

Artificial intelligence (AI) risk-management guidance emphasizes governance, traceability, monitoring, and human oversight [42–44]. Hallucination surveys show that generated text can be unsupported by or inconsistent with available evidence [45], and RAG evaluation methods have accordingly separated retrieval and generation quality [46]. Citation-generation research also demonstrates that producing a citation is not equivalent to establishing that the citation supports the claim [47]. Two recent studies support the design choices adopted here without addressing EVSE conformity. Keskin [48] found, for private-document question answering with locally adapted models, that explicit supervision on unanswerable inputs mattered more for reliable abstention than the adaptation method, which supports a typed missing-evidence state, and Oz and Keskin [49] operationalize the NIST AI Risk Management Framework for LLM and RAG deployments through verifiable artifacts such as model cards, test plans, and monitoring records, the logic followed by the pinned model and knowledge-base identifiers, audit records, and the human-approval step of this framework. A related division of labor appears in manufacturing inspection: Lee et al. [50] clustered deep embeddings of PCB defect proposals and reserved the accept-or-reject decision for a human operator at the cluster level, describing the result as an auditable path to scalable supervision; the present framework assigns the same confirmation role to human review of verdicts and citations. In a safety- and regulation-relevant workflow, a correct verdict with an invalid citation and an incorrect verdict with a valid-looking citation are both failures. Accordingly, retrieval, verdict, citation, applicability, and evidence sufficiency were evaluated separately.

## 3. Materials and Methods

### 3.1. Research Design, Questions, and Scope

The research followed the workflow in Figure 2: (i) fix the standards corpus and regulatory context; (ii) perform requirement-level cross-standard mapping; (iii) identify

device, operational, and ecosystem gaps using an EV-charging threat model; (iv) transform the analysis into an integrated requirement and evidence model; (v) implement local-LLM pre-assessment variants; and (vi) evaluate them against design-gold cases. The traceability chain was defined as Clause  $\rightarrow$  Gap  $\rightarrow$  Requirement  $\rightarrow$  Evidence  $\rightarrow$  Verdict  $\rightarrow$  Citation. Every transformation retained identifiers so that an output verdict could be traversed back to the standard version and product evidence that supported it.

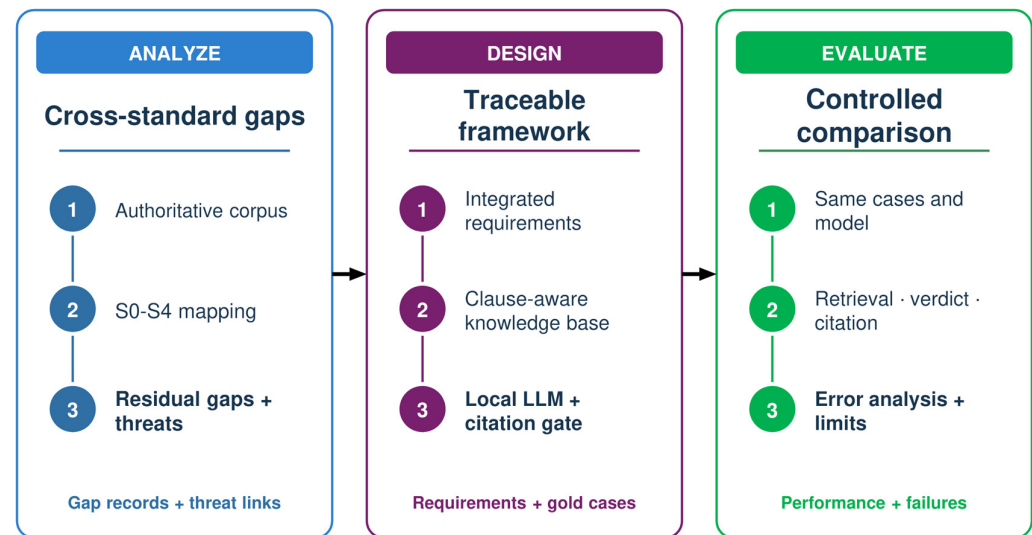


**Figure 2.** Overall research workflow.

Five research questions structured the work. RQ1 asked how strongly IEC 62443-4-2, ETSI EN 303 645, and NIST IR 8473 correspond to EN 18031 and what residual gaps remain for EV charging. RQ2 asked whether the mapping and gaps can be transformed into evidence-testable integrated requirements. RQ3 asked whether clause-aware RAG improves retrieval over fixed-length RAG. RQ4 asked whether the local-LLM variants reproduce the design-gold five-state verdicts. RQ5 asked whether citation verification and human review expose or remove unsupported definitive verdicts.

The direct assessment scope was an internet-connected EV charger with fixed hardware, firmware, communication, and option configuration. Evidence types included product scope declarations, security architecture, interface and service inventories, software bills of materials (SBOMs), configuration records, certificate and key-management procedures, update policies, logs, and test reports. Packet-level vulnerability discovery, real-time intrusion detection, dynamic distribution-grid stability analysis, and formal certification decisions were outside the scope.

### 3.2. Five-Level Cross-Standard Correspondence Mapping

The unit of mapping was the union of 44 EN 18031 requirement identifiers across Parts 1–3. For each requirement  $r$  and comparison source  $t$ , the study identified the most directly corresponding clause or clause group and assigned a strength  $S(r,t)$  from S0 to S4. Four decision factors were applied: whether the same asset, interface, and actor were addressed; whether the requirement was mandatory or recommended; whether the granularity was comparable; and whether the requirement could be verified by a specified document, configuration, or test. Table 2 defines the scale, and Figure 3 gives representative mappings.

The IEC and ETSI relationships began from the informative annexes of EN 18031 and were checked against IEC 62443-4-2:2019 and ETSI EN 303 645 V3.1.3. Because the EN annex reflects an earlier ETSI clause structure, every retained ETSI identifier was checked against V3.1.3

and updated where provisions had been split or renumbered. NIST IR 8473 was mapped independently at the CSF subcategory level. Since NIST IR 8473 is a voluntary ecosystem profile rather than a product conformance standard, semantic coverage was recorded, but its abstraction and reduced testability limited the strength assigned to general matches.

**Table 2.** Five-level correspondence scale.

| Score | Label      | Operational Definition                                                                                |
|-------|------------|-------------------------------------------------------------------------------------------------------|
| S4    | Stronger   | The comparison source is more stringent or covers an additional directly relevant security dimension. |
| S3    | Equivalent | Substantially equivalent scope, granularity, normative force, and verifiability.                      |
| S2    | Partial    | Only part of the subject, actor, strength, method, or evidence expectation corresponds.               |
| S1    | Implicit   | A high-level principle is related, but no concrete product-level requirement is provided.             |
| S0    | None       | No corresponding requirement was identified in the comparison source.                                 |

| EN 18031 requirements | IEC 62443-4-2 | ETSI EN 303 645 | NIST IR 8473 |
|-----------------------|---------------|-----------------|--------------|
| ACM-1                 | S4            | S2              | S3           |
| AUM-1                 | S4            | S2              | S4           |
| SCM-1                 | S3            | S3              | S4           |
| SUM-1                 | S2            | S4              | S3           |
| LGM-1                 | S4            | S0              | S3           |
| NMM-1                 | S3            | S0              | S4           |
| UNM-1                 | S0            | S3              | S1           |

S0: None
  S1: Implicit
  S2: Partial
  S3: Equivalent
  S4: Stronger

**Figure 3.** Illustrative strength-scale correspondence mapping.

The mapping criteria and all 132 pairwise outcomes are disclosed in Appendix A. Publishing these data permits independent review of boundary judgments and recalculation of the aggregates. The mapping was performed by one researcher. No independent expert-panel adjudication was performed, so no inter-rater statistic can be reported. Consensus procedures such as a Delphi panel remain a future validation step [51].

Let  $R$  be the 44-requirement reference set and  $T$  the comparison-source set. The strength function and three aggregate indicators were defined as follows:

$$S : R \times T \rightarrow \{0, 1, 2, 3, 4\} \quad (1)$$

$$\text{Cov}(t) = \frac{|\{r \in R : S(r, t) \geq 2\}|}{|R|} \quad (2)$$

$$\text{Cov}^*(t) = \frac{|\{r \in R : S(r, t) \geq 3\}|}{|R|} \quad (3)$$

$$\text{Score}(t) = \frac{\sum_{r \in R} S(r, t)}{4|R|} \quad (4)$$

Cov counts partial or stronger relationships; Cov\* counts equivalent or stronger relationships. Score normalizes the mean ordinal score to 0–1 and was used only as a descriptive ranking aid, not as an interval-scale measurement. Because one boundary reclassification moves Cov or Cov\* by 1/44 (2.3 percentage points) and a one-level change moves Score by 1/176 (0.006), the reported decimals indicate reproducible computation rather than equivalent measurement precision.

### 3.3. EV-Charging Threat Model and Analytical Gap Coverage

The charging system was divided into field/local, operations, and external/grid trust zones (Figure 1). Adversaries included near-field, remote-network, supply-chain, insider, and coordinated actors. Protected assets included charge and safety control, identity and payment, personal and location data, firmware, V2G transactions, and grid stability. Eight scenarios were derived using STRIDE categories and EV-specific cyber–physical consequences [52], with ISO/SAE 21434 used as a complementary reference for vehicle cybersecurity engineering concepts [53]. Table 3 summarizes the scenarios and the residual capabilities not fully supplied by the standards corpus.

**Table 3.** EV-charging threat scenarios and residual capability gaps.

| ID   | Threat Scenario                                         | Primary Consequence                                        | Residual Capability Gap                                                                       |
|------|---------------------------------------------------------|------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| TS-1 | Coordinated fleet-load attack                           | Tampering/denial of service; distribution-grid instability | Real-time DSO coordination and grid-synchronized staged recovery                              |
| TS-2 | Manipulation of bidirectional V2G flow                  | Tampering with dispatch, power quality, or settlement      | Integrated transaction authentication, settlement accountability, and grid-impact analysis    |
| TS-3 | Plug and Charge certificate theft                       | Spoofing and unauthorized charging                         | ISO 15118 certificate lifecycle management linked to theft detection and operations           |
| TS-4 | Wireless or power-line communication (PLC) interference | Denial of service and unsafe repeated retries              | Cross-physical-layer interference detection, safe stop, retry suppression, and field response |
| TS-5 | Payment or session-token replay                         | Spoofing, billing fraud, and roaming disputes              | Replay detection across session/payment/roaming and settlement reversal or dispute handling   |
| TS-6 | Firmware supply-chain compromise                        | Tampering and persistent fleet compromise                  | SBOM-to-vulnerability-feed linkage and coordinated operational blocking                       |
| TS-7 | Charging-data privacy breach                            | Information disclosure and behavioral profiling            | Cross-provider data minimization and profiling controls                                       |
| TS-8 | Combined physical–cyber intrusion                       | Tampering, privilege escalation, and service disruption    | Physical-tamper detection linked to remote operational response                               |

The scenarios build on the protocol-level literature already cited. OCPP threat studies [23–26] supply the attack primitives of TS-1, TS-5, TS-6, and TS-8 (backend impersonation, replay, malicious firmware, and denial of service), and the ISO 15118 certificate and TLS analyses [17,22] underlie TS-2 and TS-3. These protocols enter the threat model and the residual-gap analysis but were not used as mapping axes in addition to EN 18031 because they are interface specifications whose security provisions depend on the deployed profile and on backend configuration outside the manufacturer’s sole control, not assessable product requirements with applicability conditions, evidence expectations, and acceptance rules. Protocol-specific expectations therefore appear as required evidence and assessment methods of the integrated requirements (for example, the certificate-lifecycle procedures and replay tests of DEV-03 and OPS-03 in Appendix B).

For each threat  $\tau$ , the required capability set  $R(\tau)$  combined the EN 18031-side requirements directly involved in mitigation with atomic capabilities parsed from the residual-gap descriptions (the trace matrix is given in Appendix C, Table A6). A capability was counted as covered by a comparison standard only when its mapping strength was S3 or S4. The single-standard condition took the maximum coverage achieved by IEC, ETSI, or NIST for each threat. The union condition counted an S3/S4 relationship in any of the three sources. The integrated-requirement condition counted a residual capability only when an explicit

proposed requirement addressed it. The two criteria are deliberately asymmetric. External sources were credited only at S3 or S4, whereas the integrated catalogue, written to close the identified gaps, was credited on existence. So, its figures are trace-completeness values that are not comparable with the S3-threshold coverage of the external sources.

$$\text{Coverage}(c, \tau) = \frac{|\{r \in R(\tau) : \text{covered}_c(r)\}|}{|R(\tau)|} \quad (5)$$

$$\text{Coverage}_{\text{single}}(\tau) = \max_{t \in T} \text{Coverage}(t, \tau) \quad (6)$$

The integrated-requirement coverage calculated from this same trace matrix is a design-coverage check, not independent evidence of effectiveness. It shows which identified gaps were translated into requirements; it does not demonstrate that the proposed controls reduce real-world risk. This circularity was explicitly separated from the model-evaluation experiment.

### 3.4. Integrated Requirement, Product-Evidence, and Verdict Model

A correspondence map is descriptive and cannot itself decide whether a product is ready for assessment. Each strong relationship and residual gap was therefore transformed into an atomic integrated requirement (IR). An IR records a testable normative statement, distinguishes official source clauses from research-proposed gap controls, defines applicability, lists necessary product evidence, specifies an assessment method and acceptance rule, links relevant threats, and versions the complete trace. Table 4 summarizes the schema, and the 16 integrated requirements used in the evaluation are listed in Appendix B (Tables A4 and A5).

The model used three responsibility layers. Device requirements concern internal access control, communication, key storage, updates, interfaces, and resilience. Operational requirements concern vulnerability handling, support period, SBOM operation, logging, certificate lifecycle, and data deletion. Ecosystem requirements concern the charge point operator (CPO), e-mobility service provider, roaming, supplier, and DSO responsibilities. A proposed ecosystem control was never presented as if it were an EN 18031 clause. In the trace model (Figure 4) and in every report, research-proposed bases are drawn dashed, in a different color, and labeled as proposals to distinguish them from the solid authoritative clauses, so that a manufacturer cannot mistake a recommendation for a legal duty. Figure 4 shows the bidirectional trace from source clause to verdict.

The verdict model distinguishes non-applicability from missing evidence and missing evidence from verified failure. Table 5 states the five output states. For safety- or legally critical controls such as firmware signature verification, confidential-key protection, remote-management authentication, and payment functions, a document claim alone could not justify PASS without functional or test evidence.

**Table 4.** Integrated requirement schema.

| Field                  | Purpose                                | Example                                                              |
|------------------------|----------------------------------------|----------------------------------------------------------------------|
| requirement_id/version | Stable identifier and revision         | DEV-03 v1.0                                                          |
| layer/owner            | Responsibility and evidence location   | Device/manufacturer                                                  |
| statement              | Atomic, verifiable normative statement | Mutual authentication shall protect the remote-management interface. |
| source_clause          | Source, version, clause, and locator   | EN 18031-1 SCM-1; IEC 62443-4-2 CR 1.2                               |

Table 4. Cont.

| Field                    | Purpose                                                      | Example                                                 |
|--------------------------|--------------------------------------------------------------|---------------------------------------------------------|
| correspondence_strength  | Trace to S0–S4 mapping                                       | S3 (equivalent)                                         |
| gap_basis/proposal_basis | Missing dimension and justification for any proposed control | EVSE-CSMS operational linkage required for TS-1         |
| applicability            | Product condition that activates the requirement             | Remote-management interface present                     |
| required_evidence        | Documents, configuration, logs, and tests needed             | Interface list, TLS configuration, test log             |
| assessment/rule          | Method and acceptance criterion                              | Review + functional test/mutual authentication succeeds |
| threat/priority          | Risk trace and treatment priority                            | TS-3, TS-5/high                                         |
| trace/snapshot           | Model, prompt, knowledge-base, and review history            | KB-2026-07-28-a1; reviewer decision                     |

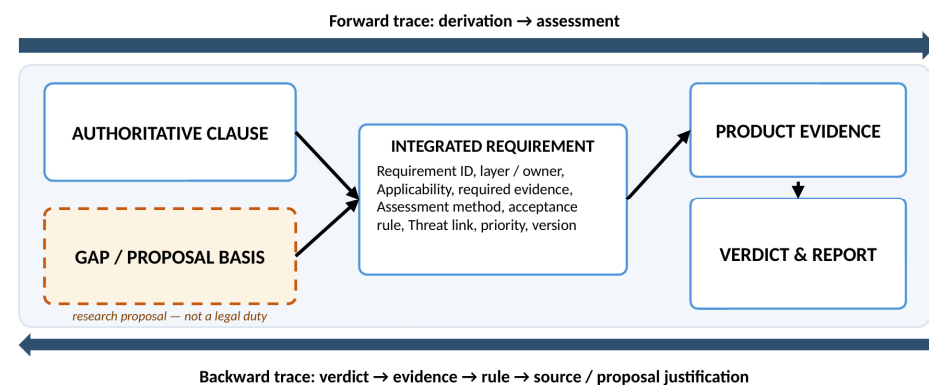


Figure 4. Trace model from standard clause to product evidence, verdict, and report.

Table 5. Five-state cybersecurity conformity pre-assessment verdict model.

| Verdict               | Condition                                                                          | Required Output                                    |
|-----------------------|------------------------------------------------------------------------------------|----------------------------------------------------|
| PASS                  | Applicable; all required evidence is present and the acceptance rule is met        | Clause citation, product evidence, and test result |
| PARTIAL               | Only part of the scope, implementation, or evidence satisfies the requirement      | Separate satisfied and missing aspects             |
| FAIL                  | Applicable; sufficient contradictory evidence or verified nonimplementation exists | Contradiction and remediation action               |
| N/A                   | The activating condition is absent from the fixed product scope                    | Scope-based non-applicability rationale            |
| INSUFFICIENT EVIDENCE | Applicability or conformity cannot be decided from available evidence              | Withhold verdict and list additional evidence      |

### 3.5. Clause-Aware Local-LLM Pre-Assessment Framework

The standards knowledge base was constructed from authoritative documents while preserving the document identifier and version, Part/Clause/Subclause hierarchy, requirement ID, normative status, scope, asset, interface, linked definitions, decision trees, tables, annexes, cross-references, page locator, and source-file hash. Regulatory metadata recorded the *Official Journal of the European Union* (OJEU) citation status, restrictions, and the transition dates discussed in Section 1. The base implementation contained 447 clause

cards; 92 cards that exceeded the 8192-token embedding limit were additionally split while retaining a parent-clause link.

Figure 5 shows the clause-aware retrieval and citation-verification pipeline, together with the five output states: N/A is decided at the applicability gate before retrieval; PASS, PARTIAL, and FAIL are confirmed only after the citation gate; a failed retrieval or citation check yields INSUFFICIENT EVIDENCE; and every state is routed to human review. Retrieval was designed as a hybrid of identifier/keyword search, concept (semantic) search, and reranking; keyword retrieval can be implemented using probabilistic relevance methods such as BM25 [54]. The evaluated configuration (Section 3.7) implemented only clause-preserving segmentation with semantic retrieval. The identifier/keyword channel and the reranker, drawn dashed in Figure 5, were not evaluated, so the retrieval gains of Section 4.3 are attributable to clause-aware segmentation alone. In the pipeline as drawn, retrieved clause cards enter the judgment prompt; condition T7 (Section 3.6) instead passes them only to the citation step, the dual-path variant that Section 5.2 takes up.

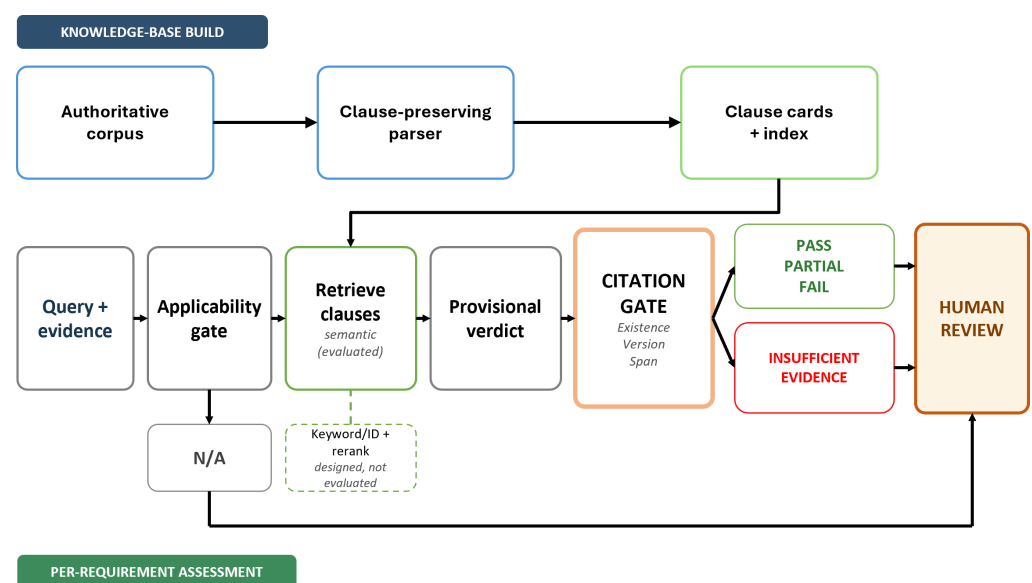


**Figure 5.** Clause-aware retrieval-augmented generation and citation-verification pipeline.

The verdict workflow, summarized in Table 6, fixed the product and firmware scope, evaluated applicability, collected product evidence, retrieved source clauses, assessed conceptual/functional/operational evidence, generated a provisional five-state result, verified citations, listed missing evidence and remediation, and routed the record to a human reviewer. Each output included the model digest, prompt version, knowledge-base snapshot, source locators, and review history. An automatic citation gate checked that each cited clause exists in the knowledge base, that its version matches the corpus version, and that the quoted span actually occurs in that clause and required at least one standard citation and one product-evidence citation for a definitive verdict. If any check failed, the definitive PASS/PARTIAL/FAIL result was withheld or downgraded to INSUFFICIENT EVIDENCE. Whether a cited clause substantively supports the claim (claim–clause entailment) requires semantic judgment and is left to human review in the workflow (Section 3.7).

Figure 6 presents the layered local architecture. The Compliance Verification Agent (CVA) performs requirement-by-requirement applicability, evidence-sufficiency, and verdict reasoning. A Gap Remediation Agent (GRA) can suggest missing evidence, tests, or corrective action without creating legal obligations. An Orchestration Agent (OA) can control sequencing, versioning, review requests, and audit logs. The contribution lies in the traceable workflow, not in the use of multiple agents. Agents were separated only where

this improved error isolation or evidence completeness, consistent with the broader multi-agent literature [55]. Local execution reduced external transmission of licensed standards and confidential product documents but was not treated as an accuracy guarantee.

**Table 6.** Requirement-level pre-assessment procedure.

| Step | Operation                                                                                      | Failure-Safe Behavior                                                        |
|------|------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| 1    | Fix product, hardware, firmware, options, and assessment scope                                 | Do not merge evidence across versions                                        |
| 2    | Evaluate the integrated requirement's applicability condition                                  | Route ambiguous scope to human review                                        |
| 3    | Retrieve and normalize required product evidence                                               | Missing evidence remains missing; it is not converted to FAIL                |
| 4    | Retrieve candidate standard clauses using the required version                                 | No out-of-corpus substitution                                                |
| 5    | Assess conceptual, completeness, sufficiency, and operational evidence                         | Record conflicts and uncertainty                                             |
| 6    | Generate a provisional five-state verdict                                                      | No definitive verdict without both clause and product evidence               |
| 7    | Check clause existence, version, and quoted span mechanically; assess claim support separately | Withhold or downgrade on a failed check; route claim support to human review |
| 8    | Generate missing-evidence and remediation actions                                              | Label research proposals separately from standard duties                     |
| 9    | Obtain human approval and write the audit record                                               | Retain pre-review and post-review outcomes                                   |

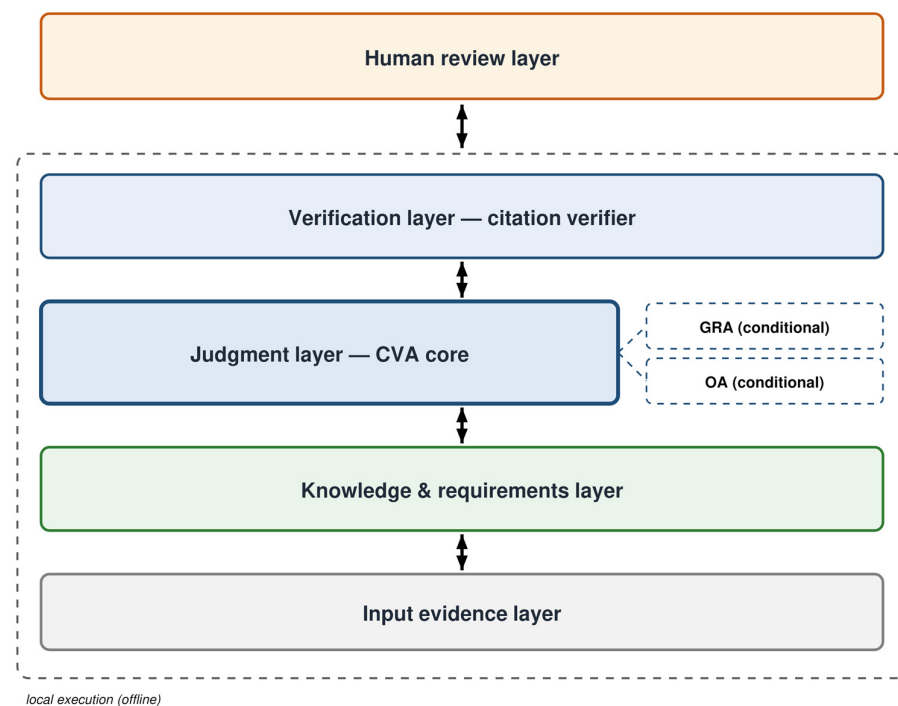


**Figure 6.** Layered local-LLM framework.

### 3.6. Evaluation Dataset and Comparison Conditions

The judgment dataset sampled 16 integrated requirements (Appendix B): six device, five operational, and five ecosystem requirements. Five cases were constructed for each requirement, one for each verdict, producing 80 balanced synthetic cases (Table 7). Each case contained a product profile, scoped functions and options, input evidence, applicability gold, verdict gold, source-clause gold, rationale, and difficulty tags. PASS cases included both design and test evidence. N/A cases removed the activating feature. FAIL cases contained affirmative contradictory evidence, and INSUFFICIENT EVIDENCE cases omitted an indispensable item. The gold label was therefore derived deterministically from case-design variables rather than assigned retrospectively. Because a classifier given the

design variables would reproduce the gold labels by construction, the accuracies reported below measure how reliably each condition recovers those variables from the evidence text, not conformity reasoning on real dossiers (Sections 4.4 and 5.4).

**Table 7.** Evaluation dataset design.

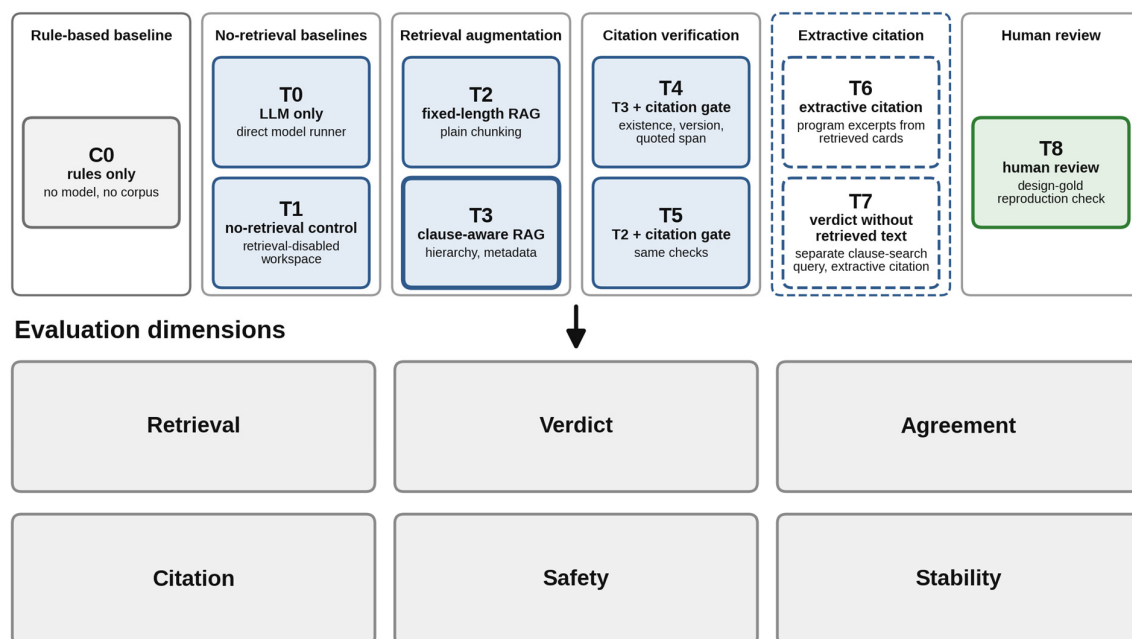
| Element                 | Configuration                                                                                   |
|-------------------------|-------------------------------------------------------------------------------------------------|
| Integrated requirements | 16: device 6, operation 5, ecosystem 5                                                          |
| Judgment cases          | 80 total; five verdict cases per requirement                                                    |
| Verdict balance         | 16 each: PASS, PARTIAL, FAIL, N/A, INSUFFICIENT EVIDENCE                                        |
| Evidence forms          | Design, configuration, SBOM, procedure, logs, and functional or penetration tests               |
| Difficulty              | Direct and indirect wording, conflicting evidence, annex/cross-reference, and version confusion |
| Gold construction       | Single deterministic label derived from controlled case-design variables                        |
| Split                   | Development 20; validation 16; holdout 44 (reported subset)                                     |
| Retrieval queries       | 24 scored questions plus two negative controls                                                  |

The cases were split and stratified by verdict into development (20), validation (16), and holdout (44) subsets. Development cases were used to draft the prompts and requirement rules and validation cases to check the drafts. Holdout cases were not used for tuning, so all verdict and citation metrics are reported on the holdout subset ( $n = 44$ , gold distribution 9/9/9/9/8) unless stated otherwise, and Appendix D (Table A7) lists the accuracies on the full set and on each subset. The split was made at the case level. The five cases of a requirement share its statement, applicability condition, required evidence, and acceptance rule, and 14 of the 16 requirements contribute cases to all three subsets, so holdout accuracy measures generalization to unseen evidence configurations of known requirements, not to unseen requirements, and a rule tuned on a development case can benefit the holdout cases of the same requirement. Retrieval evaluation used 24 scored queries with known target clauses and two unscored negative controls (a nonexistent clause identifier and an out-of-corpus question); the queries are a separate benchmark from the judgment cases, scored by the rank of the target clause, and were not used for tuning, so they are not split. Cohen's kappa [56] quantified model-gold agreement. Because the gold labels came from one researcher's deterministic case generator, Fleiss' kappa was not applicable [57]. Human review (T8) was performed by the case designer on the 20 safety- or legally critical cases (the automatic-approval exclusion set of Section 3.4), with the design gold hidden on the review card until each verdict had been recorded; this self-reported, non-blind procedure makes T8 an internal reproduction check rather than an independent benchmark.

All primary conditions used the same cases, the same base model (Qwen2.5 14B, 4-bit GGUF Q4\_K\_M), the same judgment prompt, and the same five-state output schema. Only retrieval, citation, or execution-path configuration changed, and Phi-4 14B at the same quantization was run as a cross-model diagnostic (Section 4.5). Table 8 defines the nine conditions. Every judgment prompt carried the integrated requirement statement, required evidence, and assessment method together with the product scope, product profile, and evidence (query variant P-ctx, Section 3.7), so T0 to T3 differ only in retrieval. T4 and T5 were separate runs of the T3 and T2 configurations with the citation-verification gate of Section 3.5 applied to each output before the verdict was recorded. T6 and T7 moved citation text from the language model to programmatic extraction from retrieved clause cards (query variant P-ext); T7 additionally kept retrieved text out of the judgment prompt. Its verdict was produced on the retrieval-disabled path of T1, and a separate short clause-search query (the requirement statement and required evidence) selected the cards from which the citation was extracted. Figure 7 summarizes the conditions and the evaluation dimensions applied to them.

**Table 8.** Comparison conditions.

| Condition | Configuration                                                                                 | Purpose                                                |
|-----------|-----------------------------------------------------------------------------------------------|--------------------------------------------------------|
| T0        | Local LLM without retrieval; direct model-runner path                                         | No external evidence baseline                          |
| T1        | Retrieval-disabled workspace with zero documents                                              | Path-matched no-retrieval control                      |
| T2        | Fixed-length general RAG                                                                      | Baseline retrieval augmentation                        |
| T3        | Clause-aware RAG                                                                              | Effect of hierarchy- and metadata-preserving units     |
| T4        | T3 plus citation-verification downgrading                                                     | Expose unsupported definitive T3 verdicts              |
| T5        | T2 plus the same citation-verification rule                                                   | Test whether high verdict accuracy is also traceable   |
| T6        | Model verdict plus programmatically extracted citation                                        | Separate citation production from language generation  |
| T7        | Verdict without retrieved text; a separate clause-search query feeds extractive citation only | Separate the verdict path from the citation path       |
| T8        | Human review of automated verdicts and evidence                                               | Internal reproduction and qualitative error comparison |

**Figure 7.** Experimental conditions and evaluation dimensions.

Because every gold label is a deterministic function of the case-design variables, the cases were also scored with a short rule-based classifier, C0, that uses neither a language model nor the corpus (Section 4.4): four ordered lists of regular expressions over the evidence text (7 FAIL, 7 INSUFFICIENT EVIDENCE, 6 N/A, and 6 PARTIAL markers) and, when no marker matches, the generation-role prior disclosed in the product-profile descriptions (EVC-A → PASS, EVC-B → PARTIAL, EVC-C → N/A). The rules were fixed before scoring and applied once; the rule file and scoring scripts are released with the data.

### 3.7. Metrics, Statistical Analysis, and Reproducibility

Five-class verdict performance was measured by accuracy, macro-precision, macro-recall, macro-F1, confusion matrices, and Cohen's kappa [56]. Because the holdout gold distribution (9/9/9/9/8) is nearly balanced, macro-recall almost equals accuracy and is omitted. Macro-precision assigns zero to a class with no predictions, and full confusion matrices are released with the data (error types are summarized in Section 4.3). Kappa was interpreted descriptively rather than used as a regulatory threshold; established agreement categories were used only as context [58]. Retrieval was measured by Recall@k, Precision@k, mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG@k) [59]. Citation performance was decomposed into citation existence, exact-gold-clause agreement, and unsupported-claim rate (UCR), defined below. Claim–citation support was left to

human review. Safety indicators included false PASS for known FAIL/INSUFFICIENT EVIDENCE cases and the separately reported N/A-to-PASS applicability error. Table 9 summarizes these evaluation dimensions and metrics.

**Table 9.** Evaluation dimensions and metrics.

| Dimension | Metrics                                  | Interpretation                                                    |
|-----------|------------------------------------------|-------------------------------------------------------------------|
| Retrieval | Recall@k, Precision@k, MRR, nDCG@k       | Whether target clauses occur high in the candidate list           |
| Verdict   | Accuracy, macro-P/R/F1, confusion matrix | Five-state decision performance                                   |
| Agreement | Cohen's kappa                            | Agreement with deterministic design gold beyond chance            |
| Citation  | Existence, exact clause, UCR             | Mechanical checkability; substantive support left to human review |
| Safety    | False-pass rate, N/A-to-PASS, UCR        | Dangerous approval and unsupported-claim risk                     |
| Stability | Run agreement and accuracy variance      | Sensitivity to uncontrolled runtime variation                     |

The citation metrics use different denominators, and every value is reported with its numerator and denominator. Let  $D_c$  be the definitive verdicts (PASS, PARTIAL, FAIL) of condition  $c$  on the holdout cases,  $S_c$  the multiset of standard-clause citations attached to them,  $K$  the (document, version, clause) identifiers in the knowledge-base snapshot,  $E(v)$  the number of citations of verdict  $v$  that resolve to an identifier in  $K$ ,  $C(v)$  the clause identifiers cited by  $v$ , and  $D_c^g \subseteq D_c$  the definitive verdicts on requirements whose gold source clause is a scorable standard identifier. Citation existence (8) is a citation-level rate; UCR (7) and exact-gold-clause agreement (9) are verdict-level rates. UCR counts a definitive verdict as unsupported when none of its citations exist in the corpus; it does not measure whether an existing citation is the governing clause (that is (9)) or whether the clause supports the claim, which is left to human review. UCR is undefined, not zero, when a condition leaves no definitive verdict. Under programmatic extraction (T6, T7), every attached citation exists by construction, so UCR records only whether a span was attached and (9) is the substantive metric. The gate of T4 and T5 applies a separate quotation-based test (the span exists, its version matches, and it lies in the cited clause; Section 3.5), whose rejection counts are reported in Section 4.5.

$$\text{UCR}(c) = \frac{|\{v \in D_c : E(v) = 0\}|}{|D_c|} \quad (7)$$

$$\text{Existence}(c) = \frac{|\{s \in S_c : s \in K\}|}{|S_c|} \quad (8)$$

$$\text{ExactGold}(c) = \frac{|\{v \in D_c^g : \text{gold}(v) \in C(v)\}|}{|D_c^g|} \quad (9)$$

Paired verdict differences used two-sided exact McNemar tests on the discordant pairs [60], reported as  $b$  (cases correct only under the first condition) and  $c$  (cases correct only under the second) with every  $p$ -value. With 44 cases, the smallest difference that can reach  $p < 0.05$  is six discordant cases in one direction ( $b = 6, c = 0, p = 0.0312$ ), so nonsignificance is not evidence of equivalence. Families of follow-up comparisons were adjusted by the Holm–Bonferroni procedure, and raw and adjusted  $p$ -values are both reported (Section 4.5). Macro-F1 intervals used 1000 percentile-bootstrap replicates with seed 20240501 [61]. A Wilcoxon signed-rank procedure was pre-specified for time comparison [62]; because a comparable timing baseline was outside the study design, no efficiency claim was tested. The study reported all negative and null results.

Experiments ran in the local environment of Table 10, which also records the session-isolation control (one new chat session per call on the workspace path, so that no earlier conversation is prepended to the prompt), the run windows (all Qwen2.5 runs consecu-

tively on 31 August 2026; retrieval-metric and Phi-4 runs on 3 September 2026), and the retrieval settings of each condition. The primary judgment model was Qwen2.5 14B [63] in 4-bit GGUF Q4\_K\_M quantization, with low-bit quantization treated as a possible external-validity limitation [64,65]; Phi-4 14B [66] at the same quantization provided the second-family diagnostic. The embedding model was BGE-M3 [67] (1024 dimensions) with LanceDB and cosine distance; no reranker or hybrid retrieval was enabled.

**Table 10.** Local LLM experimental environment and reproducibility records.

| Item                     | Recorded Value                                                                                                                                                                                                                                                                                                                                          |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Operating system/CPU/RAM | Windows 11 Pro/AMD Ryzen 9 7900X3D/64 GB                                                                                                                                                                                                                                                                                                                |
| GPU                      | NVIDIA GeForce RTX 4070 Ti SUPER, 16 GB VRAM (driver 566.03, CUDA 12.7)                                                                                                                                                                                                                                                                                 |
| LLM runtime              | Ollama 0.32.15 (T0 judgment calls, direct model-runner path); AnythingLLM v1.12.1 (T1 to T7 judgment calls and all workspace retrieval; T4 and T5 run the T3 and T2 configurations with the citation gate)                                                                                                                                              |
| Primary model            | qwen2.5:14b (digest 7cdf5a0187d5, 14.8B parameters, Q4_K_M, context 32,768)                                                                                                                                                                                                                                                                             |
| Second model family      | phi4:latest (digest ac896e5b8b34, 14.7B parameters, Q4_K_M, context 16,384)                                                                                                                                                                                                                                                                             |
| Embedding/vector store   | bge-m3 (digest 790764642607, 1024 dimensions)/LanceDB (cosine distance)                                                                                                                                                                                                                                                                                 |
| Generation settings      | temperature 0.0, top-p 0.9, top-k 40; seed not fixed (not exposed on all tool paths)                                                                                                                                                                                                                                                                    |
| Session isolation        | One new chat session per call on the workspace path (identifier = run tag   condition   case ID) so that no earlier conversation is prepended to the prompt; workspace history length 0                                                                                                                                                                 |
| Run window               | Qwen2.5 runs (all conditions, three repetitions) executed consecutively on 31 August 2026; retrieval-metric runs and the second model family on 3 September 2026; human review as in the initial submission                                                                                                                                             |
| Retrieval settings       | T0 none (direct)/T1 and the T7 verdict call: empty workspace/T2, T5 tool-default chunking/T3, T4, T6 447 clause cards (92 additionally split)/T7 citation path: the same cards queried with a separate clause-search query, programmatic citation extraction only/top-k 6, similarity threshold 0.2 (cosine distance); no hybrid retrieval or reranking |
| Prompts and snapshot     | One English system prompt and the judgment-query templates (released with the data); judgment-query variants P-ctx, P-ext; knowledge-base snapshot KB-2026-07-28-a1                                                                                                                                                                                     |

One system prompt was used throughout, with two judgment-query variants: P-ctx (requirement identifier, requirement statement, required evidence, assessment method, product scope, product profile, and evidence) and P-ext (P-ctx with the standard-citation field left empty for programmatic filling). All reported tables used knowledge-base snapshot KB-2026-07-28-a1. The model and embedding digests, prompt files, standard-file hashes, raw and parsed outputs, and scoring records were retained and are released with the per-case records (Supplementary Material), and primary scoring was checked by a second implementation. To assess seed and tool-path variation, every condition was run three times on all 80 cases and scored on the holdout cases.

## 4. Results

### 4.1. Cross-Standard Correspondence

Across the 44 EN 18031 requirement identifiers, NIST IR 8473 produced the broadest partial-or-stronger correspondence: Cov = 95.5% (42/44), Cov\* = 70.5% (31/44), and Score = 0.74. IEC 62443-4-2 yielded Cov = 84.1% (37/44), Cov\* = 54.5% (24/44), and Score = 0.60. ETSI EN 303 645 yielded Cov = 68.2% (30/44), Cov\* = 40.9% (18/44), and Score = 0.48 (Table 11 and Figure 8).

These rankings do not establish standard substitution. NIST IR 8473 frequently corresponds at an ecosystem or governance level rather than with a device-test requirement. IEC 62443-4-2 is comparatively strong for identification and authentication, access control,

logging, zone protection, and resource availability, but it omits consumer-facing notification and several privacy-specific applicability cases. ETSI EN 303 645 is strongest for unique passwords, vulnerability disclosure, software update, user deletion, and selected privacy practices, but it has no direct counterpart for the EN 18031 logging group and little for network monitoring or traffic control. The full row-level outcomes in Appendix A show that all three sources contain both strengths and absences.

**Table 11.** EN 18031 correspondence by comparison source.

| Comparison Source | Cov (S2–S4)   | Cov* (S3–S4)  | Score |
|-------------------|---------------|---------------|-------|
| IEC 62443-4-2     | 84.1% (37/44) | 54.5% (24/44) | 0.60  |
| ETSI EN 303 645   | 68.2% (30/44) | 40.9% (18/44) | 0.48  |
| NIST IR 8473      | 95.5% (42/44) | 70.5% (31/44) | 0.74  |

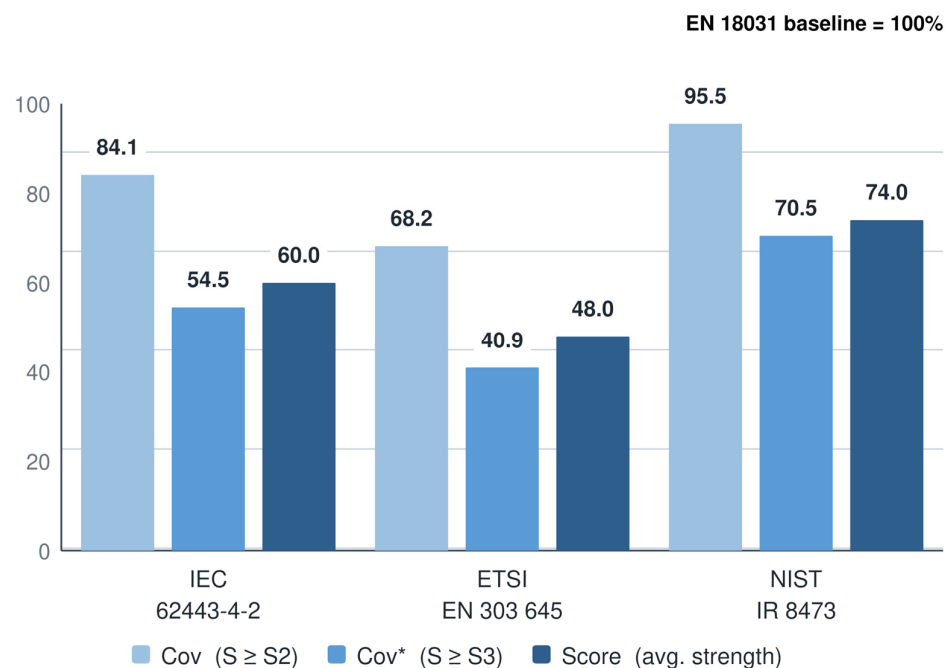


**Figure 8.** EN 18031 coverage by comparison source.

Table 11 rests on single-researcher coding of all 132 outcomes without independent adjudication (Section 3.2). Appendix A lists every assignment so that a reader can re-score any boundary judgment: one reclassification across the S2/S3 boundary moves Cov\* by 1/44 (2.3 percentage points), one across the S1/S2 boundary moves Cov by the same amount, and ETSI EN 303 645 would need six additional S3–S4 assignments to reach the Cov\* of IEC 62443-4-2, so the ranking is robust to several such changes but individual percentages are not.

#### 4.2. EV-Charging Threat Gaps

No single comparison source covered all of the equivalent or stronger capabilities required by any of the eight threat scenarios (Figure 9). Aggregated across scenario requirements, the best single source per scenario covered 70.3%, and the union of IEC, ETSI, and NIST covered 75.0% (Table 12). The integrated-requirement figures report design coverage under the existence criterion of Section 3.3, are not independent validation results, and are not comparable with the S3-threshold columns. Complementarity was visible for Plug and Charge certificate theft (63.6% single-source maximum versus 72.7% union), firmware supply-chain compromise (66.7% versus 77.8%), and charging-data privacy breach (75.0%

versus 87.5%). Other scenarios showed no increase because the same residual capability was absent across all three sources.

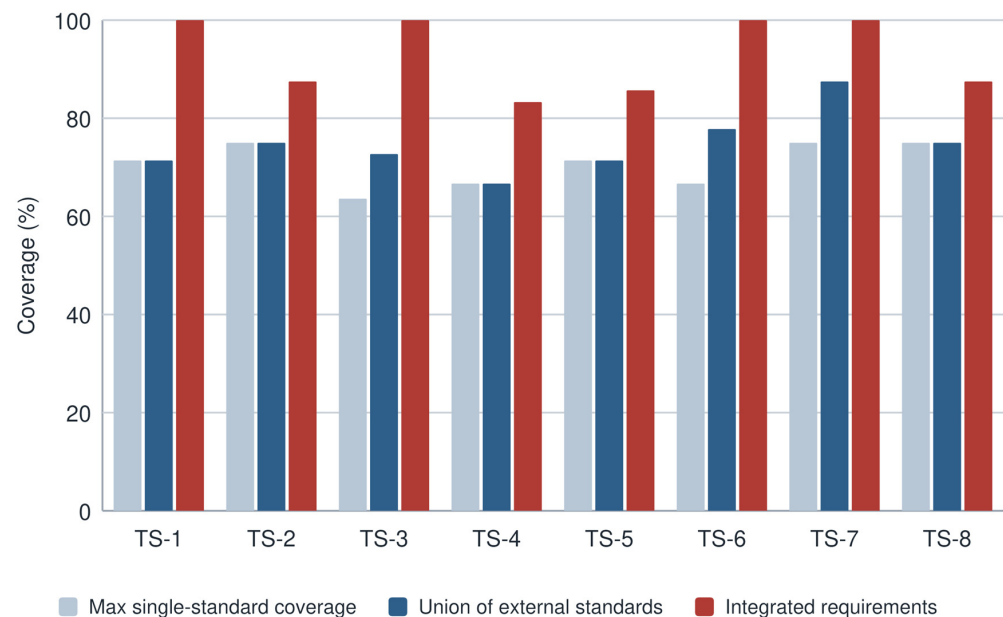


**Figure 9.** Analytical gap coverage by threat scenario.

The EN 18031 reference axis alone covered 76.6% of the defined capability set, and adding all four standards did not remove the remaining 23.4% structural gaps. Those gaps included DSO-coordinated load response, grid-impact assessment for manipulated V2G transactions, cross-physical-layer interference detection, settlement dispute linkage, and physical-tamper-to-operations response. The integrated requirement catalogue addressed most, but not all, identified gaps and reached 93.8% design coverage. As stated in Section 3.3, this value is a trace-completeness result derived from the same design matrix, not evidence of control effectiveness.

**Table 12.** Threat-scenario analytical coverage by external sources (S3 or higher) and integrated-requirement design response (%).

| Threat                                 | Best Single External Source | External-Source Union | Integrated Requirements (Design Response) |
|----------------------------------------|-----------------------------|-----------------------|-------------------------------------------|
| TS-1 Coordinated fleet load            | 71.4 (IEC/NIST)             | 71.4                  | 100.0                                     |
| TS-2 V2G-flow manipulation             | 75.0 (NIST)                 | 75.0                  | 87.5                                      |
| TS-3 Plug and Charge certificate theft | 63.6 (IEC)                  | 72.7                  | 100.0                                     |
| TS-4 Wireless/PLC interference         | 66.7 (IEC/NIST)             | 66.7                  | 83.3                                      |
| TS-5 Payment-token replay              | 71.4 (NIST)                 | 71.4                  | 85.7                                      |
| TS-6 Firmware supply chain             | 66.7 (NIST)                 | 77.8                  | 100.0                                     |
| TS-7 Charging-data privacy             | 75.0 (ETSI/NIST)            | 87.5                  | 100.0                                     |
| TS-8 Physical-cyber intrusion          | 75.0 (IEC/NIST)             | 75.0                  | 87.5                                      |
| Aggregated                             | 70.3                        | 75.0                  | 93.8                                      |

#### 4.3. Retrieval, Verdict, and Citation Results

Clause-aware segmentation substantially improved retrieval (Table 13). On the 24 scored queries, Recall@5 rose from 0.4583 for fixed-length RAG (T2) to 0.7917 for clause-aware RAG (T3), a significant gain by the exact McNemar test ( $b = 1$ ,  $c = 9$ ,  $p = 0.0215$ ), and MRR and nDCG@5 rose accordingly, although T3 placed the target clause at rank one for only 10 of the 24 queries (7 for T2). Because the evaluated configuration had no identifier/keyword channel and no reranker, the gain is attributable to clause-preserving segmentation (boundaries together with the retained hierarchy and metadata, which the

comparison does not separate), not to hybrid retrieval or reranking. The apparent retrieval scores of T0 and T1 (0.1667 and 0.1250) are direct response matching without a corpus, not a retrieval system. In the two negative controls, the no-retrieval condition produced citation-formatted entries for a nonexistent clause (XYZ-9) and for a nonexistent document, whereas both retrieval conditions stated that no such clause exists and answered the out-of-corpus question without citing the corpus. Retrieval therefore suppressed fabricated identifiers within the corpus but did not produce refusal of out-of-scope questions, which the applicability and citation gates must enforce.

**Table 13.** Retrieval, verdict, and citation results (Qwen2.5 and C0; 24 queries, 44 holdout cases).

| Metric                           | C0<br>(Rules)      | T0                 | T1                 | T2                 | T3                 | T4                 | T5                | T6                | T7                 |
|----------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|-------------------|-------------------|--------------------|
| Recall@5                         | –                  | 0.1667             | 0.1250             | 0.4583             | 0.7917             | 0.7917             | 0.4583            | 0.7917            | 0.7917             |
| Precision@5                      | –                  | 0.0333             | 0.0250             | 0.0917             | 0.1583             | 0.1583             | 0.0917            | 0.1583            | 0.1583             |
| MRR                              | –                  | 0.1667             | 0.1250             | 0.3604             | 0.5625             | 0.5625             | 0.3604            | 0.5625            | 0.5625             |
| nDCG@5                           | –                  | 0.1667             | 0.1250             | 0.3681             | 0.6202             | 0.6202             | 0.3681            | 0.6202            | 0.6202             |
| Accuracy                         | 0.6591             | 0.8636             | 0.8864             | 0.7727             | 0.7727             | 0.6364             | 0.4545            | 0.7500            | 0.9318             |
| Macro-precision                  | 0.5325             | 0.8788             | 0.8978             | 0.8347             | 0.8444             | 0.7929             | 0.7085            | 0.8051            | 0.9436             |
| Macro-F1                         | 0.5607             | 0.8670             | 0.8883             | 0.7584             | 0.7684             | 0.6449             | 0.4545            | 0.7508            | 0.9327             |
| Macro-F1 95% bootstrap CI        | [0.459, 0.655]     | [0.754, 0.957]     | [0.778, 0.976]     | [0.622, 0.876]     | [0.627, 0.885]     | [0.478, 0.769]     | [0.285, 0.593]    | [0.604, 0.868]    | [0.853, 1.000]     |
| Cohen’s kappa                    | 0.5725             | 0.8296             | 0.8580             | 0.7169             | 0.7170             | 0.5490             | 0.3270            | 0.6887            | 0.9147             |
| N/A-to-PASS errors (of 9)        | 0                  | 2                  | 2                  | 3                  | 2                  | 2                  | 2                 | 2                 | 2                  |
| Definitive verdicts ( <i>n</i> ) | 31                 | 28                 | 28                 | 27                 | 25                 | 20                 | 13                | 24                | 29                 |
| Citation existence               | –                  | 0.0000/<br>(0/29)  | 0.0000/<br>(0/29)  | 0.4643/<br>(13/28) | 0.6333/<br>(19/30) | 0.7500/<br>(18/24) | 0.6923/<br>(9/13) | 1.0000 *          | 1.0000 *           |
| Exact-gold-clause agreement      | 0.0000/<br>(0/13)  | 0.0000/<br>(0/10)  | 0.0000/<br>(0/10)  | 0.2000/<br>(2/10)  | 0.3750/<br>(3/8)   | 0.2857/<br>(2/7)   | 0.1250/<br>(1/8)  | 0.3333/<br>(3/9)  | 0.3636/<br>(4/11)  |
| Unsupported-claim rate (UCR)     | 1.0000/<br>(31/31) | 1.0000/<br>(28/28) | 1.0000/<br>(28/28) | 0.5185/<br>(14/27) | 0.3600/<br>(9/25)  | 0.2500/<br>(5/20)  | 0.3077/<br>(4/13) | 0.0000/<br>(0/24) | 0.3448/<br>(10/29) |

Retrieval rows use the 24 scored queries. T0 and T1 use no retrieval, and their rows score the clauses the model named without a corpus (T1 on the empty-workspace path); T4, T5, T6, and the T7 citation path share the retrieval configuration of T3, T2, T3, and T3. Verdict and citation rows use the 44 holdout cases (P-ctx for T0–T5, P-ext for T6 and T7); parentheses give numerator/denominator as defined in Section 3.7. The T4 and T5 rows score the retained definitive verdicts; the gate tests quoted spans whereas (8) tests identifiers, so retained citations with an incomplete identifier (NIST IR 8473 subcategories without the function prefix) fail (8). C0 attaches no citation. \* 1.0000 by construction. Macro-precision assigns zero to a class with no predictions.

Retrieved clause text did not improve five-state verdicts (Table 13; error types by condition in Table 14). The no-retrieval conditions reached 0.8636 (T0) and 0.8864 (T1), against 0.7727 for both RAG conditions (T2, T3). None of these paired differences were significant at  $n = 44$  (Table 15, family A), but their direction was the same in the development and validation subsets (Appendix D). The loss under retrieval was concentrated in abstention. T2 and T3 returned INSUFFICIENT EVIDENCE for 14 and 15 of the 44 cases against 9 for T0 and T1 (Table 14). With the requirement statement present in every prompt, retrieved clause text led the model to withhold verdicts rather than to decide them correctly, and the highest accuracy, 0.9318, came from T7, whose judgment prompt contained no retrieved text at all (Section 4.5).

Citations under the two retrieval conditions mostly existed but rarely named the governing clause (existence 0.4643 and 0.6333, exact-gold-clause agreement 0.2000 and 0.3750 for T2 and T3), whereas the no-retrieval conditions never cited a clause that exists in the corpus (existence 0.0000, UCR 1.0000). A model writing citations from memory produced citation-formatted entries naming other standards or nonexistent clauses. With the citation gate applied (T4, T5), definitive verdicts whose citations failed the checks were withheld as INSUFFICIENT EVIDENCE, and accuracy fell to 0.6364 and 0.4545 (Section 4.5).

**Table 14.** Error-type distribution on the 44 holdout cases (Qwen2.5).

| Condition | Errors (of 44) | Most Frequent Error Types (Gold → Output)                                                                                                      | N/A → PASS (of 9) | Retreats from<br>PASS/PARTIAL/FAIL to<br>INSUFFICIENT<br>EVIDENCE |
|-----------|----------------|------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|-------------------------------------------------------------------|
| T0        | 6              | N/A → PASS 2; PASS → PARTIAL 2; PARTIAL → INSUFFICIENT EVIDENCE 1; FAIL → PARTIAL 1                                                            | 2                 | 1                                                                 |
| T1        | 5              | N/A → PASS 2; PARTIAL → INSUFFICIENT EVIDENCE 1; FAIL → PARTIAL 1; PASS → PARTIAL 1                                                            | 2                 | 1                                                                 |
| T2        | 10             | N/A → INSUFFICIENT EVIDENCE 3; N/A → PASS 3; PARTIAL → INSUFFICIENT EVIDENCE 2; FAIL → INSUFFICIENT EVIDENCE 1; PASS → PARTIAL 1               | 3                 | 3                                                                 |
| T3        | 10             | N/A → INSUFFICIENT EVIDENCE 3; FAIL → INSUFFICIENT EVIDENCE 3; N/A → PASS 2; PARTIAL → INSUFFICIENT EVIDENCE 1; PASS → PARTIAL 1               | 2                 | 4                                                                 |
| T4        | 16             | N/A → INSUFFICIENT EVIDENCE 4; PARTIAL → INSUFFICIENT EVIDENCE 4; FAIL → INSUFFICIENT EVIDENCE 3; N/A → PASS 2; PASS → INSUFFICIENT EVIDENCE 2 | 2                 | 9                                                                 |
| T5        | 24             | PARTIAL → INSUFFICIENT EVIDENCE 7; N/A → INSUFFICIENT EVIDENCE 5; PASS → INSUFFICIENT EVIDENCE 5; FAIL → INSUFFICIENT EVIDENCE 4               | 2                 | 16                                                                |
| T6        | 11             | PARTIAL → INSUFFICIENT EVIDENCE 4; N/A → INSUFFICIENT EVIDENCE 2; N/A → PASS 2; PASS → PARTIAL 2                                               | 2                 | 5                                                                 |
| T7        | 3              | N/A → PASS 2; FAIL → PARTIAL 1                                                                                                                 | 2                 | 0                                                                 |

Error types are listed in descending frequency; ties at the last listed rank are included. The last column counts gold PASS, PARTIAL, or FAIL cases returned as INSUFFICIENT EVIDENCE. T4 and T5 are the T3 and T2 configurations with the citation gate.

**Table 15.** Exact McNemar comparisons on the 44 holdout cases with Holm–Bonferroni adjustment.

| Family                                                 | Comparison<br>(First vs. Second) | Accuracy          | b  | c  | p (Raw) | p (Holm) |
|--------------------------------------------------------|----------------------------------|-------------------|----|----|---------|----------|
| A. Model conditions T0–T7                              | T0 vs. T1                        | 0.8636 vs. 0.8864 | 0  | 1  | 1.0000  | 1.0000   |
|                                                        | T0 vs. T2                        | 0.8636 vs. 0.7727 | 5  | 1  | 0.2188  | 1.0000   |
|                                                        | T0 vs. T3                        | 0.8636 vs. 0.7727 | 5  | 1  | 0.2188  | 1.0000   |
|                                                        | T1 vs. T2                        | 0.8864 vs. 0.7727 | 5  | 0  | 0.0625  | 0.6875   |
|                                                        | T1 vs. T3                        | 0.8864 vs. 0.7727 | 6  | 1  | 0.1250  | 1.0000   |
|                                                        | T2 vs. T3                        | 0.7727 vs. 0.7727 | 4  | 4  | 1.0000  | 1.0000   |
|                                                        | T2 vs. T6                        | 0.7727 vs. 0.7500 | 3  | 2  | 1.0000  | 1.0000   |
|                                                        | T3 vs. T6                        | 0.7727 vs. 0.7500 | 4  | 3  | 1.0000  | 1.0000   |
|                                                        | T6 vs. T7                        | 0.7500 vs. 0.9318 | 0  | 8  | 0.0078  | 0.1016   |
|                                                        | T0 vs. T6                        | 0.8636 vs. 0.7500 | 5  | 0  | 0.0625  | 0.6875   |
| B. Effect of citation verification on accuracy         | T0 vs. T7                        | 0.8636 vs. 0.9318 | 0  | 3  | 0.2500  | 1.0000   |
|                                                        | T1 vs. T6                        | 0.8864 vs. 0.7500 | 6  | 0  | 0.0312  | 0.3750   |
|                                                        | T1 vs. T7                        | 0.8864 vs. 0.9318 | 0  | 2  | 0.5000  | 1.0000   |
|                                                        | T3 vs. T4                        | 0.7727 vs. 0.6364 | 6  | 0  | 0.0312  | 0.0312   |
| C. Verification coverage (definitive verdict retained) | T2 vs. T5                        | 0.7727 vs. 0.4545 | 14 | 0  | 0.0001  | 0.0002   |
|                                                        | T4 vs. T5                        | 0.4545 vs. 0.2955 | 9  | 2  | 0.0654  | 0.0654   |
| D. Rule-based baseline C0 vs. model conditions         | C0 vs. T0                        | 0.6591 vs. 0.8636 | 5  | 14 | 0.0636  | 0.2543   |
|                                                        | C0 vs. T1                        | 0.6591 vs. 0.8864 | 4  | 14 | 0.0309  | 0.1544   |
|                                                        | C0 vs. T2                        | 0.6591 vs. 0.7727 | 9  | 14 | 0.4049  | 1.0000   |
|                                                        | C0 vs. T3                        | 0.6591 vs. 0.7727 | 7  | 12 | 0.3593  | 1.0000   |
|                                                        | C0 vs. T6                        | 0.6591 vs. 0.7500 | 10 | 14 | 0.5413  | 1.0000   |
|                                                        | C0 vs. T7                        | 0.6591 vs. 0.9318 | 2  | 14 | 0.0042  | 0.0251   |

b: Cases correct only under the first condition; c: cases correct only under the second (family C: cases retaining a definitive verdict only under the first or the second condition). All tests use the 44 holdout cases and were recomputed from the released per-case records of the session-isolated runs. Family A was adjusted as one thirteen-test family, B as one two-test family, and D as one six-test family; C is a single test. Comparisons within the Phi-4 family and between the model families are given in Section 4.5.

The most frequent error was not a false PASS on an applicable failing case but an applicability error. PASS was returned for a non-applicable product on 2 of the 9 hold-out N/A cases in every model condition except T2 (3), T7 included (Table 14). Under retrieval, retreats to INSUFFICIENT EVIDENCE, including for non-applicable cases, were the dominant error. The conventional false-pass metric, whose numerator counts FAIL and INSUFFICIENT EVIDENCE cases but not N/A, was zero in every model condition, so N/A-to-PASS and abstention had to be reported separately to see the dominant unsafe behavior and where retrieval lost accuracy.

#### 4.4. Diagnostic Experiments: Execution Path and Rule-Based Baseline

A path-matched no-retrieval control (T1) routed the same model and query through an empty workspace with retrieval disabled (every call logged zero retrieved sources). Its holdout accuracy matched T0 on the direct runner path ( $b = 0$ ,  $c = 1$ ), so the execution path does not explain the differences between conditions, and the same control is the most direct test of whether clause retrieval helps the verdict: T1 was above T3 by six cases against one and above T2 by five against zero (Table 15, family A). Retrieving clause text into the judgment context never improved verdict accuracy in these data, which motivates the separation of the verdict path from the citation path that T7 implements (Section 4.5) and Section 5.2 generalizes.

The rule-based classifier C0 (Section 3.6) reached 0.6591 on the holdout cases with no N/A-to-PASS error (Table 13). A text marker matched in only 5 of the 44 cases. The other 39 were decided by the profile prior, so the score is essentially what the product profile alone yields, and its errors were of kinds the prior cannot see: FAIL returned as PARTIAL (7; C0 returned no FAIL at all), INSUFFICIENT EVIDENCE returned as PASS (6 of the 8 cases with missing evidence), and FAIL returned as INSUFFICIENT EVIDENCE (2). Every model condition was above C0 by 4 to 12 net cases, but only the T7 difference survived Holm adjustment within family D of Table 15 (adjusted  $p = 0.0251$ ). Two conclusions follow. First, about two-thirds of the accuracy is available without reading the evidence, so the accuracy of the model conditions should be read as recovery of the design variables from the evidence text (Section 3.6), and the 0.09 to 0.27 that they add is the part that requires reading the evidence. Second, C0 never converted a non-applicable product into PASS, which supports the deterministic applicability gate of Section 5.2, but it converted missing evidence into PASS. So, such a gate must be paired with a typed missing-evidence state.

#### 4.5. Extractive Citations, Verification, and Cross-Model Checks

Programmatic citation extraction changed the citation failure mode (Table 13). T6 reached 0.7500 accuracy, not significantly below T3 or T2, while UCR fell to 0.0000 because every one of its 24 definitive verdicts carried a span extracted from a retrieved clause card. Citation existence was thus solved structurally, but only three of the nine scorable definitive verdicts selected the exact gold clause, no better than the model-written citations of T3 (three of eight), and claim–clause support was not automated. T6 eliminates citations to nonexistent or wrong-version clauses. But ensuring that the cited clause governs the requirement still requires better clause selection (reranking, identifier lookup), and whether it supports the verdict remains a human review task.

T7 kept retrieved text out of the judgment prompt and used a separate clause-search query only to select the cards from which the citation was extracted. Its verdict accuracy, 0.9318, was the highest of any condition: above T6 by eight cases against none ( $p = 0.0078$ ; Holm-adjusted  $p = 0.1016$  within family A), above T2 and T3 by seven against none, and above T0 and T1 by three and two against none. Its three errors were two N/A-to-PASS cases, and one FAIL returned as PARTIAL; it never retreated to INSUFFICIENT EVIDENCE

(Table 14). Exact-gold-clause agreement was 0.3636 (4/11), but the citation path was incomplete. For 10 of the 29 definitive verdicts, all on five requirements (DEV-05, OPS-02, ECO-01, ECO-02, and ECO-03), the clause search returned no card with a normative sentence to extract, so no citation was attached (UCR 0.3448). Keeping retrieved text out of the judgment prompt thus removed the abstention that retrieval induced in T2, T3, and T6. The focused clause-search query improved clause selection at no verdict cost, and extraction coverage became the bottleneck. These results support a dual-path architecture: full requirement and acceptance context for the verdict, a focused query only for the citation path, and reconciliation through the requirement-to-clause trace matrix.

The Phi-4 14B diagnostic reproduced the citation result but not the verdict ordering. Its holdout accuracy ranged from 0.7727 (T0, T3, T6, T7) to 0.8409 (T1, T2). No paired difference between the model families or within the Phi-4 family was significant (largest T7:  $b = 10$ ,  $c = 3$  in favor of Qwen2.5,  $p = 0.0923$ , Holm-adjusted  $p = 0.7383$ ), so the advantage of T7 over T6 ( $b = 7$ ,  $c = 7$ ) did not recur. Phi-4's dominant error was the opposite of Qwen2.5's. It rarely converted a non-applicable product into PASS but converted INSUFFICIENT EVIDENCE and N/A cases into PARTIAL; it over-decided rather than abstained. Citation existence again depended on the mechanism rather than on the model: with model-written citations, the UCR was 1.0000 for T0 and T1 and 0.2121 and 0.1212 for T2 and T3, with exact-gold-clause agreement at most 0.0667; with extractive citations UCR was 0.0000 for T6 and 0.3438 for T7, with exact-gold-clause agreement 0.2500 and 0.3846, and the extraction-coverage gap of T7 recurred (11 of 32 definitive verdicts without a citation). Per-condition Phi-4 metrics are released with the data, and its accuracies by subset are listed in Appendix D.

Applying the citation gate to the model-written conditions removed part of the definitive verdicts (Table 13). T4, the T3 configuration with the gate, rejected the citations of 9 of the 44 outputs, including 1 N/A verdict because the gate as implemented also demanded a corpus-valid citation for N/A, so definitive-verdict coverage fell from 0.5682 to 0.4545 and accuracy from 0.7727 to 0.6364 ( $b = 6$ ,  $c = 0$ ,  $p = 0.0312$ ). Seventeen of the 20 retained definitive verdicts matched the gold, against 22 of 25 before the gate. T5, the T2 configuration with the gate, rejected 20 outputs; coverage fell from 0.6136 to 0.2955 and accuracy to 0.4545 ( $b = 14$ ,  $c = 0$ ,  $p = 0.0001$ ), with 10 of 13 retained verdicts matching the gold (23 of 27 before). In the second family, the gate removed 20 and 21 of the 44 outputs, leaving accuracy 0.5227 for both, with 13 of 17 and 11 of 14 retained verdicts matching the gold. The gate therefore did not consistently raise the precision of the retained verdicts, and its demotion of N/A verdicts is a side effect that the deterministic applicability gate of Section 5.2 avoids, since non-applicability is a scope decision rather than a claim requiring a clause citation. Verification exposes unsupported outputs but cannot replace a citation-production mechanism that supplies valid evidence for the verdicts it removes. Table 15 collects all paired comparisons with discordant counts and Holm-adjusted  $p$ -values.

#### 4.6. Human Review (T8)

Human review (T8) was applied to the 20 safety- or legally critical cases for which Section 3.4 excludes automatic approval (the five cases of each of DEV-01, DEV-02, DEV-03, and OPS-03), 11 of which lie in the holdout subset. The reviewer, who was the case designer, opened the design gold only after recording each verdict and reproduced the gold in all 20 cases. This agreement is a ceiling produced by the review design, not an accuracy score, and reviewer time was not measured. On the same 11 holdout cases, the model conditions without the gate agreed with the gold in 9 or 10 cases (Qwen2.5) and 8 to 11 cases (Phi-4), and C0 in 6. In the primary family, every disagreement concerned a non-applicable product (one N/A case returned as PASS under every condition, a second

as INSUFFICIENT EVIDENCE under the retrieval conditions). The second family added PARTIAL returned for non-applicable or under-evidenced cases, and in T7, FAIL returned for partially fulfilled or under-evidenced requirements. Human review is therefore the corrective layer for these error types, and the only layer that assesses claim–clause support. But the design does not show that it is faster or more accurate than an unassisted assessor; a blind review by assessors who did not design the cases remains future work (Section 5.4).

#### 4.7. Repeatability

Three complete 80-case runs of every condition were performed consecutively on the same runtime (Table 16). All eight conditions returned identical verdicts on all 44 holdout cases in all three runs (accuracy standard deviation 0.0000, pairwise run agreement 1.0000), so the reported ordering does not depend on the run. Stability did not guarantee correctness, since T5 was as repeatable as T7 at half its accuracy.

**Table 16.** Three-run stability (44 holdout cases per run).

| Metric                            | T0      | T1      | T2      | T3      | T4      | T5      | T6      | T7      |
|-----------------------------------|---------|---------|---------|---------|---------|---------|---------|---------|
| Run accuracies                    | 0.8636/ | 0.8864/ | 0.7727/ | 0.7727/ | 0.6364/ | 0.4545/ | 0.7500/ | 0.9318/ |
|                                   | 0.8636/ | 0.8864/ | 0.7727/ | 0.7727/ | 0.6364/ | 0.4545/ | 0.7500/ | 0.9318/ |
|                                   | 0.8636  | 0.8864  | 0.7727  | 0.7727  | 0.6364  | 0.4545  | 0.7500  | 0.9318  |
| Mean accuracy                     | 0.8636  | 0.8864  | 0.7727  | 0.7727  | 0.6364  | 0.4545  | 0.7500  | 0.9318  |
| Accuracy standard deviation       | 0.0000  | 0.0000  | 0.0000  | 0.0000  | 0.0000  | 0.0000  | 0.0000  | 0.0000  |
| Mean macro-F1                     | 0.8670  | 0.8883  | 0.7584  | 0.7684  | 0.6449  | 0.4545  | 0.7508  | 0.9327  |
| Mean Cohen’s kappa                | 0.8296  | 0.8580  | 0.7169  | 0.7170  | 0.5490  | 0.3270  | 0.6887  | 0.9147  |
| Cases identical in all three runs | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  |
| Pairwise run agreement            | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  | 1.0000  |
| Majority-vote accuracy            | 0.8636  | 0.8864  | 0.7727  | 0.7727  | 0.6364  | 0.4545  | 0.7500  | 0.9318  |

Three complete session-isolated 80-case runs of every condition, executed consecutively on the runtime of Table 10 and scored on the 44 holdout cases. Macro-F1 and kappa are averaged over the three runs; majority-vote accuracy scores the most frequent verdict of the three runs against the gold.

## 5. Discussion

### 5.1. Answers to the Research Questions

For RQ1, the correspondence and threat-gap analysis showed that the standards are complementary rather than substitutable. NIST IR 8473 has the widest thematic coverage but is often too abstract for a device verdict. IEC 62443-4-2 is strongest in industrial component hardening and availability. ETSI EN 303 645 is strongest in consumer-IoT lifecycle and user-facing controls, and EN 18031 provides the assessment-oriented regulatory reference axis. The 23.4% capability gap remaining after combining the four sources is concentrated in EV-specific cross-organizational and cyber–physical coordination rather than in conventional access control or cryptography.

For RQ2, the integrated requirement model and trace chain showed that conversion is feasible when official clauses and research proposals are separated, one requirement has one assessment purpose, applicability is explicit, and the acceptance rule names the required product evidence. The resulting catalogue is not a new standard; it is an audit structure that makes the origin and status of every criterion visible.

For RQ3, clause-aware segmentation increased Recall@5 from 0.4583 to 0.7917 ( $p = 0.0215$ ) and nDCG@5 from 0.3681 to 0.6202, consistent with the intuition behind RAG that retrieval units should preserve the semantic boundaries of the source [38,39]. The target clause reached the first rank in only 10 of the 24 queries, so clause preservation alone does not solve candidate ordering. The gain is attributable to clause-aware segmentation alone; the hybrid identifier/keyword channel and the reranker of Section 3.5 remain unevaluated.

RQ4 was supported only in part, and not by retrieval. With the requirement statement, required evidence, and assessment method in every prompt, the no-retrieval conditions reproduced the design gold in 0.86 to 0.89 of the holdout cases; retrieved clause text in the judgment context lowered accuracy to 0.77 through abstention, and the highest accuracy (0.93) came from T7, which uses retrieval only for citations. The differences were not individually significant at  $n = 44$ , but their direction held in every case subset, under a path-matched control, and in three repetitions. Retrieved clause text in the judgment context should therefore be treated as a source of abstention rather than as a benefit, and verdict performance was limited by what entered the judgment prompt, not by retrieval recall. Because the rule-based classifier C0 reached 0.66 on the same cases almost entirely from the profile prior, the additional accuracy of the model conditions reflects reading the evidence statements, which, as Section 3.6 notes, is recovery of the design variables rather than reasoning against the standards, and only for T7 was that addition significant after adjustment.

RQ5 was partially supported. Citation verification exposed unsupported definitive verdicts, but by reducing coverage (0.5682 to 0.4545 for T4) rather than by producing traceable answers. It did not raise the precision of the retained verdicts (17 of 20, against 22 of 25 before the gate) and could not supply evidence for the verdicts it removed. Programmatic extraction changed what can go wrong rather than how well the system cites. Every extracted citation exists by construction, so the UCR of 0 in T6 is a property of the mechanism, while the measured quantities, clause selection and claim support, remained weak (the exact gold clause in 3 of 9 scorable verdicts for T6 and 4 of 11 for T7, no more often than the model-written citations of T3, and claim support not automated). Human review of claim–clause support therefore remains mandatory, and “traceable” in this study means mechanically checkable, not verified as correct.

## 5.2. Design Implications for Traceable Local LLMs

The results separate three functions that are often collapsed into one RAG prompt. First, the verdict path needs the complete requirement statement, applicability condition, required evidence, and acceptance rule, and it was most accurate when no retrieved clause text entered the judgment prompt (T7, T1, and T0 above T2, T3, and T6). Second, the citation path needs a focused clause-retrieval query, version filter, and programmatic span extraction. Third, the verification path needs deterministic checks for clause existence, version, and quoted span, together with a human check of claim support for which no deterministic proxy was available. A single mixed prompt serves neither function well. Retrieved clause text in it lowered verdict accuracy, and a model left to write its own citations filled the field from prior knowledge (Section 4.3).

A dual-path architecture follows, and T7 is its first implementation. The verdict path receives the full assessment context and no retrieved clause text. The citation path retrieves with a focused clause-search query derived from the requirement and its source-clause trace, and the program attaches an exact source span. The two paths join only after a span is attached. If none is found, or it is version-mismatched or does not support the claim, the definitive verdict is withheld, which in T7 would have withheld 10 of 29 definitive verdicts, so extraction coverage of the citation path is the next design target. The language model may explain the evidence and propose remediation, but it does not author a clause number or quotation.

Applicability deserves a dedicated deterministic gate before model reasoning. N/A-to-PASS occurred in every model condition of the primary family (two or three of the nine non-applicable holdout cases), was the most frequent error of the no-retrieval conditions and of T7, and was invisible to a conventional false-pass definition, whereas C0,

which decides applicability from the product profile, never made it (Section 4.4). Feature flags and scope evidence should activate or deactivate requirements before the LLM sees them. Missing evidence likewise needs a typed state rather than free text. Otherwise, a model tends to convert absence into FAIL or to invent an operational test, and a profile prior converts it into PASS, as C0 did in six of the eight holdout cases with missing evidence.

Local execution functions primarily as a governance control. It reduces external disclosure, enables model and knowledge-base pinning, and supports reproducible audit logs. It does not prevent hallucination, model bias, or unstable tool behavior. Locality, clause preservation, citation extraction, verification, and human approval solve different problems and must all be designed explicitly.

### 5.3. Implications for EV-Charger Manufacturers and Assessors

The implications in this section are prospective. They follow from the design and from synthetic cases with deterministic gold labels, and none has been tested on manufacturer dossiers or laboratory intake documents. So, the framework is a design-validation proof of concept whose external validity on real, anonymized documents remains to be established.

For manufacturers, the intended value is evidence readiness. A report organized by requirement can identify missing architecture descriptions, configuration captures, negative test cases, certificate-lifecycle records, SBOM-to-vulnerability links, and version mismatches before formal submission. The system can also distinguish an official harmonized-standard requirement from a supplementary standard and a research-proposed ecosystem control, reducing the risk that a recommendation is misrepresented as a legal duty.

For testing and conformity-assessment organizations, the trace model is intended to provide a reproducible intake structure. A reviewer can inspect the selected standard version, applicability decision, exact source clause, submitted product evidence, acceptance criterion, automated rationale, verification outcome, and any human override. The report remains advisory. Accredited personnel retain responsibility for sampling, test adequacy, legal interpretation, and final conformity conclusions, and, on the present evidence, for confirming that each cited clause is the governing one and supports the claim.

For standards engineering, the threat-gap results show where conventional product controls end. Coordinated load response, V2G settlement accountability, cross-provider data governance, supply-chain fleet action, and grid-synchronized recovery depend on ecosystem interfaces. Product standards can identify evidence hooks and safe local behavior, but cross-organization responsibilities may require profiles, implementation agreements, or regulatory guidance in addition to device clauses.

### 5.4. Limitations and Validity Threats

Construct validity is limited by single-researcher mapping and threat-to-requirement tracing. Publishing the criteria and all 132 outcomes makes the analysis auditable but does not create inter-rater reliability (Section 4.1 gives the sensitivity of the aggregates to single reclassifications). Independent standards and conformity experts should re-adjudicate the mapping, applicability rules, and exact-clause gold set. The integrated-requirement coverage is calculated from the design's own trace matrix under an existence criterion and must not be interpreted as effectiveness validation or compared with the S3-threshold coverage of the external standards.

Internal validity is affected by different execution paths for T0 and the tool-mediated conditions and by the inability to fix a seed on every path; the path-matched control T1 ( $b = 0$ ,  $c = 1$  against T0) and the three identical runs reduced but did not eliminate this concern. Reporting on 44 holdout cases limits statistical power: two-thirds of the comparisons in Table 15 had fewer than 12 discordant pairs, the smallest detectable one-sided

difference is six cases, and only three of the differences discussed (the accuracy cost of the gate for T3 and for T2, and the margin of T7 over C0) survived Holm adjustment, none of them cross-model. Nonsignificance is therefore not proof of equivalence, and the consistent direction of the differences across conditions and case subsets carries more weight than any single test.

Conclusion validity is limited by deterministic synthetic gold labels created by the research team. This construction is appropriate for isolating controlled evidence differences but is not a substitute for independent human ground truth. Because each label is a function of the case-design variables, the rule-based baseline C0 recovered 0.66 of the holdout labels from the product profile alone (Section 4.4), so the reported accuracies measure recovery of the design variables from the evidence text and should not be read as conformity reasoning against a real dossier. All verdict and citation metrics are reported on the holdout cases because the development and validation cases were consulted during prompt and rule design (Section 3.6). Holdout accuracy lies below development accuracy for 10 of the 16 model conditions, by at most 0.16 (Appendix D), and, because requirements are shared across the split, it does not test generalization to unseen requirements. T8 used internal, non-blind review and was therefore treated qualitatively; a blind review would give assessors who did not design the cases the same review cards without the gold section and report their verdicts, time per case, and agreement with the gold and with each other (Cohen's kappa) alongside the automated conditions. Programmatically attached citations have existence equal to one by construction; their quality must be judged by exact-clause agreement and entailment rather than existence.

External validity is limited to 80 balanced synthetic cases (44 reported), 16 requirements, 24 retrieval queries biased toward EN 18031, and two 14B model families in 4-bit quantized form; real manufacturer documents are longer, noisier, version-inconsistent, and less balanced, and the framework has not been applied to any of them or to laboratory intake documents. Quantization effects were not isolated, and processing time was not evaluated. Exact-gold-clause agreement remained at most 0.38, and the most accurate configuration (T7) attached no citation to 10 of its 29 definitive holdout verdicts, so the study establishes design requirements and failure modes, not production readiness.

Future work should include independent expert adjudication with weighted agreement reporting, blind human review by assessors who did not design the cases, anonymized real-document evaluation, hybrid identifier/semantic retrieval with reranking, additional languages and model sizes, direct analysis of the 92 additionally split clause cards, and an assessor-controlled pilot. The highest priority is improving exact-clause selection without removing the requirement context that supports verdict accuracy.

## 6. Conclusions

This study connected EV-charging cybersecurity standards analysis to an evidence-oriented local-LLM pre-assessment workflow and evaluated it as a design-validation proof of concept. A five-level mapping, performed by one researcher and published in full for audit, showed that EN 18031, IEC 62443-4-2, ETSI EN 303 645, and NIST IR 8473 provide complementary but non-substitutable coverage. Their union still leaves EV-specific operational and ecosystem gaps, which were translated into traceable integrated requirements while remaining explicitly separate from official legal obligations.

The experiments, on 80 synthetic cases with deterministic design-gold labels and reported on the 44 holdout cases, yielded a qualified result. Clause-aware segmentation improved retrieval substantially (Recall@5 from 0.46 to 0.79), but retrieved clause text in the judgment prompt did not improve verdict accuracy. The no-retrieval conditions (0.86 and 0.89) exceeded fixed-length and clause-aware RAG (0.77 each), which lost accuracy

to abstention, and the most accurate condition (0.93) kept retrieved text out of the judgment prompt and used retrieval only to extract citations. A rule-based classifier with no model and no corpus reached 0.66 on the same cases, almost entirely from the disclosed product profile, so the measured accuracy demonstrates reading of evidence statements against a label scheme, rather than conformity reasoning against the standards. Citation reliability was a separate problem. Without retrieval, model-generated citations never existed in the corpus. With retrieval, they mostly existed but named the governing clause in at most 3 of 8 scorable verdicts; programmatic extraction guarantees that a cited clause exists but not that it is the relevant one (3 of 9 and 4 of 11), that a span is found at all (none for 10 of the 29 definitive T7 verdicts), or that it supports the claim, which remains a human-review task. Citation verification removed unsupported verdicts at the cost of coverage without recovering evidence that had never been produced.

A defensible pre-assessment architecture requires separate verdict and citation paths, deterministic applicability and evidence-sufficiency gates, programmatic source-span extraction, mechanical citation verification, and documented human approval of claim–clause support. Under these constraints, a local LLM can help organize evidence, identify gaps, and explain preliminary findings. Because the mapping was single-researcher coded, the gold labels were synthetic, and no real dossier was assessed, these conclusions are design requirements established in a controlled setting rather than evidence of readiness. Within that scope, the framework offers manufacturers and assessment bodies a traceable starting point for evidence preparation and pre-assessment, one that complements accredited cybersecurity testing and formal conformity assessment rather than replacing them.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/wevj17100507/s1>.

**Author Contributions:** Conceptualization, J.-h.K. and Y.-S.B.; methodology, J.-h.K.; software, J.-h.K.; validation, J.-h.K. and S.-h.L.; formal analysis, J.-h.K.; investigation, J.-h.K.; resources, J.-h.K.; data curation, J.-h.K.; writing–original draft preparation, J.-h.K.; writing–review and editing, S.-h.L. and Y.-S.B.; visualization, J.-h.K.; supervision, Y.-S.B.; project administration, S.-h.L.; funding acquisition, S.-h.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Ministry of Trade, Industry and Energy (MOTIE, Republic of Korea) through the Technology Innovation Program, grant number RS-2024-00443436.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| Abbreviation | Meaning                           |
|--------------|-----------------------------------|
| AI           | Artificial intelligence           |
| API          | Application programming interface |
| BOM          | Bill of materials                 |
| CPO          | Charge point operator             |
| CPU          | Central processing unit           |
| CRA          | Cyber Resilience Act              |

|        |                                                                                                         |
|--------|---------------------------------------------------------------------------------------------------------|
| CRL    | Certificate revocation list                                                                             |
| CSF    | Cybersecurity Framework                                                                                 |
| CSMS   | Charging station management system                                                                      |
| CVA    | Compliance Verification Agent                                                                           |
| CVE    | Common Vulnerabilities and Exposures                                                                    |
| DoS    | Denial of service                                                                                       |
| DSO    | Distribution system operator                                                                            |
| EV     | Electric vehicle                                                                                        |
| EVSE   | Electric vehicle supply equipment                                                                       |
| GGUF   | GGUF model file format                                                                                  |
| GRA    | Gap Remediation Agent                                                                                   |
| IACS   | Industrial automation and control systems                                                               |
| IoT    | Internet of Things                                                                                      |
| IR     | Integrated requirement                                                                                  |
| LLM    | Large language model                                                                                    |
| MRR    | Mean reciprocal rank                                                                                    |
| N/A    | Not applicable                                                                                          |
| nDCG   | Normalized discounted cumulative gain                                                                   |
| OA     | Orchestration Agent                                                                                     |
| OCPP   | Open Charge Point Protocol                                                                              |
| OCSP   | Online Certificate Status Protocol                                                                      |
| OJEU   | Official Journal of the European Union                                                                  |
| PKI    | Public key infrastructure                                                                               |
| PLC    | Power-line communication                                                                                |
| RACI   | Responsible, Accountable, Consulted, and Informed                                                       |
| RAG    | Retrieval-augmented generation                                                                          |
| RAM    | Random-access memory                                                                                    |
| RED    | Radio Equipment Directive                                                                               |
| RF     | Radio frequency                                                                                         |
| SBOM   | Software bill of materials                                                                              |
| SLA    | Service-level agreement                                                                                 |
| STRIDE | Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, and Elevation of privilege |
| TLS    | Transport Layer Security                                                                                |
| UCR    | Unsupported-claim rate                                                                                  |
| UI     | User interface                                                                                          |
| V2G    | Vehicle to grid                                                                                         |
| VEX    | Vulnerability Exploitability eXchange                                                                   |
| VRAM   | Video random-access memory                                                                              |

## Appendix A. Complete Cross-Standard Correspondence Matrix

Tables A1–A3 disclose every strength assignment used to calculate Table 11. EN parts are shown in Part 1/Part 2/Part 3 order; O indicates that the requirement is present in the corresponding EN 18031 part and X that it is not. X in a comparison-source column indicates that no corresponding requirement was identified. Clause groups separated by a slash or range were assessed together; NIST CSF 1.1 subcategory identifiers are written in full (function.category-number) so that each locator is unambiguous. The mappings record correspondence, not automatic conformity equivalence.

**Table A1.** EN 18031 correspondence with IEC 62443-4-2:2019.

| EN Req. | EN Parts | IEC 62443-4-2 Clause(s) | S |
|---------|----------|-------------------------|---|
| ACM-1   | O/O/O    | CR 1.1–1.14, 2.1–2.3    | 4 |
| ACM-2   | O/O/O    | CR 1.1–1.14, 2.1–2.3    | 4 |
| ACM-3   | X/O/X    | X                       | 0 |
| ACM-4   | X/O/X    | X                       | 0 |
| ACM-5   | X/O/X    | X                       | 0 |
| ACM-6   | X/O/X    | X                       | 0 |
| AUM-1   | O/O/O    | CR 1.1–1.14             | 4 |
| AUM-2   | O/O/O    | CR 1.1–1.14             | 4 |
| AUM-3   | O/O/O    | CR 1.5, 1.10            | 3 |
| AUM-4   | O/O/O    | CR 1.5                  | 2 |
| AUM-5   | O/O/O    | CR 1.7                  | 2 |
| AUM-6   | O/O/O    | CR 1.7, 1.11            | 3 |
| SUM-1   | O/O/O    | CR 3.10                 | 2 |
| SUM-2   | O/O/O    | CR 3.10                 | 2 |
| SUM-3   | O/O/O    | CR 3.10                 | 2 |
| SSM-1   | O/O/O    | CR 3.1, 4.1             | 3 |
| SSM-2   | O/O/O    | CR 3.1                  | 3 |
| SSM-3   | O/O/O    | CR 4.1                  | 3 |
| SCM-1   | O/O/O    | CR 3.1, 3.8, 4.1        | 3 |
| SCM-2   | O/O/O    | CR 3.1, 3.8, 4.1        | 3 |
| SCM-3   | O/O/O    | CR 4.1                  | 2 |
| SCM-4   | O/O/O    | CR 3.1, 3.8             | 2 |
| LGM-1   | X/O/O    | CR 2.8, 2.9             | 4 |
| LGM-2   | X/O/O    | CR 2.8, 2.9, 2.10       | 4 |
| LGM-3   | X/O/O    | CR 2.9, 2.10            | 3 |
| LGM-4   | X/O/O    | CR 2.11                 | 3 |
| DLM-1   | X/O/X    | CR 4.2                  | 2 |
| UNM-1   | X/O/X    | X                       | 0 |
| UNM-2   | X/O/X    | X                       | 0 |
| RLM-1   | O/X/X    | CR 7.1, 7.2, 7.4        | 4 |
| NMM-1   | O/X/X    | CR 5.2, 6.1, 6.2        | 3 |
| TCM-1   | O/X/X    | CR 5.2                  | 3 |
| CCK-1   | O/O/O    | CR 1.9, 1.14, 4.3       | 3 |
| CCK-2   | O/O/O    | CR 4.3                  | 3 |
| CCK-3   | O/O/O    | CR 4.3                  | 2 |
| GEC-1   | O/O/O    | X                       | 0 |
| GEC-2   | O/O/O    | CR 5.2, 7.6, 7.7        | 3 |
| GEC-3   | O/O/O    | CR 2.1, 5.2, 7.6        | 3 |
| GEC-4   | O/O/O    | CR 7.6                  | 2 |
| GEC-5   | O/O/O    | CR 7.7                  | 3 |
| GEC-6   | O/O/O    | CR 3.5                  | 3 |
| GEC-7   | X/O/X    | CR 7.6                  | 2 |
| GEC-8   | X/X/O    | CR 7.6                  | 2 |
| CRY-1   | O/O/O    | CR 4.3                  | 2 |

**Table A2.** EN 18031 correspondence with ETSI EN 303 645 V3.1.3.

| EN Req. | EN Parts | ETSI EN 303 645 Provision(s) | S |
|---------|----------|------------------------------|---|
| ACM-1   | O/O/O    | 5.5-4/5                      | 2 |
| ACM-2   | O/O/O    | 5.5-5, 5.6-7                 | 2 |
| ACM-3   | X/O/X    | X                            | 0 |
| ACM-4   | X/O/X    | X                            | 0 |
| ACM-5   | X/O/X    | 5.5-5                        | 2 |
| ACM-6   | X/O/X    | 5.5-5                        | 2 |

Table A2. Cont.

| EN Req. | EN Parts | ETSI EN 303 645 Provision(s)                      | S |
|---------|----------|---------------------------------------------------|---|
| AUM-1   | O/O/O    | 5.5-4/5                                           | 2 |
| AUM-2   | O/O/O    | 5.1-3                                             | 2 |
| AUM-3   | O/O/O    | X                                                 | 0 |
| AUM-4   | O/O/O    | 5.1-4                                             | 2 |
| AUM-5   | O/O/O    | 5.1-1/2/3                                         | 4 |
| AUM-6   | O/O/O    | 5.1-5                                             | 3 |
| SUM-1   | O/O/O    | 5.3-1/2, 5.3-15A/B                                | 4 |
| SUM-2   | O/O/O    | 5.3-9/10                                          | 4 |
| SUM-3   | O/O/O    | 5.3-3/5, 5.3-4A/B, 5.3-6A/B                       | 4 |
| SSM-1   | O/O/O    | 5.4-1, 5.6-3                                      | 3 |
| SSM-2   | O/O/O    | 5.4-1/2                                           | 3 |
| SSM-3   | O/O/O    | 5.4-1                                             | 2 |
| SCM-1   | O/O/O    | P1,3:5.5-6/7; P2:+5.8-1/2                         | 3 |
| SCM-2   | O/O/O    | X                                                 | 0 |
| SCM-3   | O/O/O    | P1,3:5.5-6/7; P2:+5.8-1/2                         | 3 |
| SCM-4   | O/O/O    | 5.5-1                                             | 3 |
| LGM-1   | X/O/O    | X                                                 | 0 |
| LGM-2   | X/O/O    | X                                                 | 0 |
| LGM-3   | X/O/O    | X                                                 | 0 |
| LGM-4   | X/O/O    | X                                                 | 0 |
| DLM-1   | X/O/X    | 5.9-1, 5.11-1                                     | 4 |
| UNM-1   | X/O/X    | 6-1                                               | 3 |
| UNM-2   | X/O/X    | 6-1                                               | 3 |
| RLM-1   | O/X/X    | 5.9-1                                             | 2 |
| NMM-1   | O/X/X    | X                                                 | 0 |
| TCM-1   | O/X/X    | X                                                 | 0 |
| CCK-1   | O/O/O    | X                                                 | 0 |
| CCK-2   | O/O/O    | 5.1-3                                             | 2 |
| CCK-3   | O/O/O    | 5.1-1, 5.4-4                                      | 2 |
| GEC-1   | O/O/O    | 5.2-1/2/3(indirect)                               | 1 |
| GEC-2   | O/O/O    | 5.6-1/5                                           | 3 |
| GEC-3   | O/O/O    | X                                                 | 0 |
| GEC-4   | O/O/O    | X                                                 | 0 |
| GEC-5   | O/O/O    | 5.6-1/3                                           | 3 |
| GEC-6   | O/O/O    | 5.13-1A/B                                         | 3 |
| GEC-7   | X/O/X    | 5.8-3                                             | 3 |
| GEC-8   | X/X/O    | 5.7-1/2                                           | 2 |
| CRY-1   | O/O/O    | P1:5.1-3,5.3-7,5.5-1/2/3;<br>P2:+5.8-1; P3:+5.8-2 | 4 |

Table A3. EN 18031 correspondence with NIST IR 8473.

| EN Req. | EN Parts | NIST CSF Subcategory Reference(s) | S |
|---------|----------|-----------------------------------|---|
| ACM-1   | O/O/O    | PR.AC-1, PR.AC-3, PR.AC-4         | 3 |
| ACM-2   | O/O/O    | PR.AC-1, PR.AC-3, PR.AC-4         | 4 |
| ACM-3   | X/O/X    | PR.AC-4                           | 2 |
| ACM-4   | X/O/X    | PR.AC-4, PR.DS-5                  | 2 |
| ACM-5   | X/O/X    | PR.AC-1, PR.AC-4                  | 2 |
| ACM-6   | X/O/X    | PR.AC-4, PR.DS-5                  | 2 |
| AUM-1   | O/O/O    | PR.AC-1, PR.AC-6, PR.AC-7         | 4 |
| AUM-2   | O/O/O    | PR.AC-6, PR.AC-7                  | 4 |
| AUM-3   | O/O/O    | PR.AC-1, PR.AC-6, PR.AC-7         | 3 |
| AUM-4   | O/O/O    | PR.AC-1, PR.AC-6, PR.AC-7         | 3 |
| AUM-5   | O/O/O    | PR.AC-1, PR.AC-6, PR.AC-7         | 3 |

**Table A3.** *Cont.*

| EN Req. | EN Parts | NIST CSF Subcategory Reference(s)            | S |
|---------|----------|----------------------------------------------|---|
| AUM-6   | O/O/O    | PR.AC-1, PR.AC-7, DE.CM-7                    | 3 |
| SUM-1   | O/O/O    | ID.AM-2, PR.IP-3, PR.IP-12, PR.MA-1          | 3 |
| SUM-2   | O/O/O    | PR.DS-6, PR.IP-3                             | 4 |
| SUM-3   | O/O/O    | ID.AM-2, PR.IP-3, PR.IP-12, PR.MA-1          | 3 |
| SSM-1   | O/O/O    | PR.AC-4, PR.DS-1, PR.DS-5                    | 3 |
| SSM-2   | O/O/O    | PR.DS-1, PR.DS-6                             | 3 |
| SSM-3   | O/O/O    | PR.DS-1, PR.DS-5                             | 3 |
| SCM-1   | O/O/O    | ID.AM-3, PR.DS-2, PR.PT-4                    | 4 |
| SCM-2   | O/O/O    | PR.DS-2, PR.DS-6, PR.PT-4                    | 4 |
| SCM-3   | O/O/O    | PR.DS-2, PR.DS-5, PR.PT-4                    | 4 |
| SCM-4   | O/O/O    | PR.DS-2, PR.PT-4                             | 3 |
| LGM-1   | X/O/O    | PR.PT-1, DE.AE-3                             | 3 |
| LGM-2   | X/O/O    | PR.DS-1, PR.PT-1                             | 3 |
| LGM-3   | X/O/O    | PR.DS-4, PR.PT-1                             | 2 |
| LGM-4   | X/O/O    | PR.PT-1, DE.AE-3                             | 2 |
| DLM-1   | X/O/X    | PR.DS-3, PR.IP-6                             | 3 |
| UNM-1   | X/O/X    | ID.GV-3, RS.CO-2, RS.CO-3                    | 1 |
| UNM-2   | X/O/X    | ID.GV-3, RS.CO-3                             | 1 |
| RLM-1   | O/X/X    | PR.DS-4, PR.PT-4, PR.PT-5                    | 4 |
| NMM-1   | O/X/X    | DE.AE-1, DE.AE-5, DE.CM-1                    | 4 |
| TCM-1   | O/X/X    | PR.AC-5, PR.PT-3, PR.PT-4                    | 4 |
| CCK-1   | O/O/O    | PR.DS-1, PR.DS-2, PR.PT-4                    | 2 |
| CCK-2   | O/O/O    | PR.DS-1, PR.DS-2, PR.PT-4                    | 2 |
| CCK-3   | O/O/O    | PR.AC-1, PR.DS-1, PR.DS-2, PR.PT-4           | 2 |
| GEC-1   | O/O/O    | ID.RA-1, ID.RA-2, ID.RA-5, PR.IP-12, DE.CM-8 | 4 |
| GEC-2   | O/O/O    | PR.AC-5, PR.IP-1, PR.PT-3, PR.PT-4           | 4 |
| GEC-3   | O/O/O    | PR.AC-4, PR.IP-1, PR.IP-3, PR.PT-3           | 3 |
| GEC-4   | O/O/O    | ID.AM-3, ID.AM-4, PR.IP-1                    | 3 |
| GEC-5   | O/O/O    | PR.IP-1, PR.PT-3                             | 3 |
| GEC-6   | O/O/O    | PR.DS-6, PR.IP-2, DE.CM-4                    | 2 |
| GEC-7   | X/O/X    | ID.AM-3, ID.GV-3, PR.DS-5                    | 2 |
| GEC-8   | X/X/O    | PR.DS-6, PR.DS-8, PR.PT-4                    | 4 |
| CRY-1   | O/O/O    | PR.DS-1, PR.DS-2, PR.DS-6, PR.PT-4           | 3 |

## Appendix B. Integrated EV-Charger Security Requirement Catalogue

Tables A4 and A5 present the 16 integrated requirements and their principal evidence and assessment methods, respectively. The basis column distinguishes official standard mechanisms from residual-gap proposals. Before a real assessment, each source clause and applicability condition must be rechecked against the designated product, jurisdiction, and standard version.

**Table A4.** Integrated requirement statements and trace bases.

| ID     | Layer  | Requirement Statement                                                                                     | Basis/Gap                                  | Threat |
|--------|--------|-----------------------------------------------------------------------------------------------------------|--------------------------------------------|--------|
| DEV-01 | Device | Verify boot and firmware integrity and retain a safe state after verification failure                     | SUM/GEC; IEC system integrity              | TS-6   |
| DEV-02 | Device | Protect confidential keys from unauthorized extraction or modification                                    | SSM/CCK                                    | TS-3/5 |
| DEV-03 | Device | Apply mutual authentication and replay protection to remote management and charging-backend communication | AUM/SCM; IEC identification/authentication | TS-3/5 |
| DEV-04 | Device | Remove or disable unnecessary services and externally accessible interfaces                               | GEC; ETSI 5.6                              | TS-8   |

**Table A4.** *Cont.*

| ID     | Layer     | Requirement Statement                                                                               | Basis/Gap                              | Threat |
|--------|-----------|-----------------------------------------------------------------------------------------------------|----------------------------------------|--------|
| DEV-05 | Device    | Prioritize safety-critical functions during communication failure and overload                      | RLM/TCM; IEC resource availability     | TS-1/4 |
| DEV-06 | Device    | Control physically accessible diagnostic interfaces and provide tamper evidence                     | IEC CR 2.13; residual gap              | TS-8   |
| OPS-01 | Operation | Document and operate a vulnerability-disclosure policy and security-support period                  | ETSI 5.2/5.3; NIST governance          | TS-6   |
| OPS-02 | Operation | Link SBOMs to vulnerability feeds and trace affected products                                       | NIST supply-chain risk; residual gap   | TS-6   |
| OPS-03 | Operation | Manage issuance, replacement, revocation, and expiry of Plug and Charge and management certificates | ISO 15118; NIST; residual gap          | TS-3   |
| OPS-04 | Operation | Apply trusted time, protection, and retention controls to security-event logs                       | LGM; IEC CR 2.11                       | TS-3/6 |
| OPS-05 | Operation | Minimize personal data and provide a user-data deletion process                                     | EN 18031-2; ETSI                       | TS-7   |
| ECO-01 | Ecosystem | Connect authentication, authorization, and settlement actors for V2G power transactions             | Partial NIST/ISO 15118; gap            | TS-2   |
| ECO-02 | Ecosystem | Detect coordinated-load anomalies and respond in stages with the DSO                                | Partial NIST; gap                      | TS-1   |
| ECO-03 | Ecosystem | Predetermine incident-information-sharing, responsibility, and contact structures                   | NIST response communication/governance | TS-1/6 |
| ECO-04 | Ecosystem | Operate supplier-security requirements, component traceability, and change approval                 | NIST supply chain; IEC                 | TS-6   |
| ECO-05 | Ecosystem | Perform staged, grid-synchronized recovery after a large-scale disruption                           | NIST recovery; residual gap            | TS-1/2 |

**Table A5.** Principal evidence and assessment method for integrated requirements.

| ID     | Principal Evidence                                        | Assessment Method                           |
|--------|-----------------------------------------------------------|---------------------------------------------|
| DEV-01 | Signature policy, boot log, test report                   | Tampered-package functional test            |
| DEV-02 | Key-storage design, permissions, extraction-test evidence | Document + configuration + penetration test |
| DEV-03 | Certificates, TLS configuration, packet logs              | Mutual authentication and replay test       |
| DEV-04 | Port, service, and interface inventory                    | Scanning + configuration review             |
| DEV-05 | State model, rate limits, test results                    | DoS and network-loss test                   |
| DEV-06 | Seals, port controls, and logs                            | Field inspection + access test              |
| OPS-01 | Policy, contact, SLA, and handling records                | Document review + sampled case review       |
| OPS-02 | SBOM, VEX, and CVE handling records                       | Traceability sampling                       |
| OPS-03 | PKI procedure, CRL/OCSP, and logs                         | Expiry and revocation scenario test         |
| OPS-04 | Log schema, time source, and access permissions           | Tamper and clock-error test                 |
| OPS-05 | Data map, deletion UI/API, and logs                       | Deletion-completeness test                  |
| ECO-01 | Role model, transaction logs, and settlement rules        | Cross-system scenario                       |
| ECO-02 | Aggregate metrics, alerting, and mitigation procedure     | Simulated coordinated event                 |
| ECO-03 | RACI, contact tree, and exercise records                  | Tabletop exercise                           |
| ECO-04 | Supplier assessment, BOM, and change records              | Supply-chain document sampling              |
| ECO-05 | Recovery priorities, DSO agreement, and exercise records  | Recovery scenario exercise                  |

## Appendix C. Threat-to-Requirement Trace Matrix

Table A6 is the trace used to calculate Table 12. The standard-side set contains the EN 18031 requirements directly involved in each threat. Residual gaps are atomic capabilities derived from the gap analysis. The final column records whether the integrated catalogue contains an explicit response; it does not claim independent effectiveness.

**Table A6.** Threat-to-requirement trace and residual-gap response.

| Threat | Standard-Side Requirement Set              | Residual Capability Gap                                                            | Integrated Response              |
|--------|--------------------------------------------|------------------------------------------------------------------------------------|----------------------------------|
| TS-1   | ACM-1, AUM-1, NMM-1, RLM-1, TCM-1          | (1) DSO-linked real-time load control;<br>(2) grid-synchronized recovery           | Both: device/ecosystem           |
| TS-2   | ACM-1, AUM-1, AUM-2, CRY-1, LGM-1, SCM-1   | (1) V2G transaction authentication/settlement; (2) integrated grid-impact analysis | (1) addressed; (2) open          |
| TS-3   | AUM-1/2/3, CCK-1/2/3, CRY-1, SCM-1, SSM-1  | (1) ISO 15118 certificate lifecycle; (2) theft detection linked to operations      | Both: operation/ecosystem        |
| TS-4   | NMM-1, RLM-1, SCM-1, TCM-1                 | (1) RF/PLC interference detection; (2) safe stop and retry suppression             | (2) addressed; (1) open          |
| TS-5   | AUM-1/2, CRY-1, SCM-1, SSM-1               | (1) session/payment-backend replay detection; (2) reversal/dispute linkage         | (1) addressed; (2) open          |
| TS-6   | CCK-1, CRY-1, GEC-1, SSM-1, SUM-1/2/3      | (1) SBOM-vulnerability-feed linkage; (2) operational blocking policy               | Both: device/operation/ecosystem |
| TS-7   | ACM-1, CRY-1, DLM-1, SCM-1, SSM-1/2, UNM-1 | Cross-provider minimization and profiling prevention                               | Addressed: operation             |
| TS-8   | ACM-1, GEC-2, LGM-1, NMM-1, SCM-1, SSM-1   | (1) physical-tamper detection; (2) linkage to remote response                      | (2) addressed; (1) open          |

## Appendix D. Verdict Accuracy by Case Subset

Table A7 lists, for every reported condition, the verdict accuracy on all 80 cases and on the development, validation, and holdout subsets of Section 3.6; the holdout column is the primary result reported in Table 13.

**Table A7.** Verdict accuracy by case subset for every reported condition.

| Configuration                | Condition | All ( <i>n</i> = 80) | Development ( <i>n</i> = 20) | Validation ( <i>n</i> = 16) | Holdout ( <i>n</i> = 44) |
|------------------------------|-----------|----------------------|------------------------------|-----------------------------|--------------------------|
| Qwen2.5 14B (primary family) | T0        | 0.9000               | 0.9000                       | 1.0000                      | 0.8636                   |
|                              | T1        | 0.9125               | 0.9000                       | 1.0000                      | 0.8864                   |
|                              | T2        | 0.8375               | 0.9000                       | 0.9375                      | 0.7727                   |
|                              | T3        | 0.8125               | 0.8000                       | 0.9375                      | 0.7727                   |
|                              | T4        | 0.7125               | 0.8000                       | 0.8125                      | 0.6364                   |
|                              | T5        | 0.5000               | 0.5000                       | 0.6250                      | 0.4545                   |
|                              | T6        | 0.8000               | 0.8000                       | 0.9375                      | 0.7500                   |
| Phi-4 14B (second family)    | T7        | 0.9250               | 0.8500                       | 1.0000                      | 0.9318                   |
|                              | T0        | 0.7625               | 0.7500                       | 0.7500                      | 0.7727                   |
|                              | T1        | 0.8500               | 0.8500                       | 0.8750                      | 0.8409                   |
|                              | T2        | 0.8625               | 0.8500                       | 0.9375                      | 0.8409                   |
|                              | T3        | 0.7750               | 0.7000                       | 0.8750                      | 0.7727                   |
|                              | T4        | 0.4625               | 0.3000                       | 0.5000                      | 0.5227                   |
|                              | T5        | 0.5375               | 0.5000                       | 0.6250                      | 0.5227                   |
| Rule-based baseline          | T6        | 0.8125               | 0.8000                       | 0.9375                      | 0.7727                   |
|                              | T7        | 0.8000               | 0.7500                       | 0.9375                      | 0.7727                   |
|                              | C0        | 0.6500               | 0.6500                       | 0.6250                      | 0.6591                   |

T4 and T5 are the T3 and T2 configurations with the citation-verification gate. C0 is the rule-based baseline of Section 4.4. Holdout values are the figures reported in the main text.

## References

1. International Energy Agency. Global EV Outlook 2025. Available online: <https://www.iea.org/reports/global-ev-outlook-2025> (accessed on 5 August 2026).
2. Guerrero-Silva, J.A.; Romero-Gelvez, J.I.; Aristizábal, A.J.; Zapata, S. Optimization and Trends in EV Charging Infrastructure: A PCA-Based Systematic Review. *World Electr. Veh. J.* **2025**, *16*, 345. [[CrossRef](#)]

3. Acharya, S.; Dvorkin, Y.; Pandžić, H.; Karri, R. Cybersecurity of Smart Electric Vehicle Charging: A Power Grid Perspective. *IEEE Access* **2020**, *8*, 214434–214453. [[CrossRef](#)]
4. Ronanki, D.; Karneddi, H. Electric Vehicle Charging Infrastructure: Review, Cyber Security Considerations, Potential Impacts, Countermeasures, and Future Trends. *IEEE J. Emerg. Sel. Top. Power Electron.* **2024**, *12*, 242–256. [[CrossRef](#)]
5. European Parliament and Council. Directive 2014/53/EU of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment. *Off. J. Eur. Union* **2014**, *L153*, 62–106.
6. European Commission. Commission Delegated Regulation (EU) 2022/30 of 29 October 2021 supplementing Directive 2014/53/EU with regard to Article 3(3)(d), (e), and (f). *Off. J. Eur. Union* **2022**, *L7*, 6–10.
7. European Commission. Commission Delegated Regulation (EU) 2023/2444 of 20 July 2023 amending Delegated Regulation (EU) 2022/30 as regards the date of application of the essential requirements for radio equipment and correcting that Regulation. *Off. J. Eur. Union* **2023**, *L153*, 62.
8. EN 18031-1; Common Security Requirements for Radio Equipment—Part 1: Internet Connected Radio Equipment. CEN (European Committee for Standardization): Brussels, Belgium, 2024.
9. EN 18031-2; Common Security Requirements for Radio Equipment—Part 2: Radio Equipment Processing Data, Namely Internet Connected Radio Equipment, Childcare Radio Equipment, Toys Radio Equipment and Wearable Radio Equipment. CEN (European Committee for Standardization): Brussels, Belgium, 2024.
10. EN 18031-3; Common Security Requirements for Radio Equipment—Part 3: Internet Connected Radio Equipment Processing Virtual Money or Monetary Value. CEN (European Committee for Standardization): Brussels, Belgium, 2024.
11. European Commission. Commission Implementing Decision (EU) 2025/138 of 28 January 2025 amending Implementing Decision (EU) 2022/2191 as regards harmonised standards in support of the essential requirements of Directive 2014/53/EU of the European Parliament and of the Council that relate to cybersecurity, for the categories and classes of radio equipment specified in Delegated Regulation (EU) 2022/30. *Off. J. Eur. Union* **2025**, *2025/138*.
12. European Parliament and Council. Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). *Off. J. Eur. Union* **2024**, *2024/2847*.
13. European Commission. Commission Delegated Regulation (EU) 2026/339 of 16 February 2026 repealing Delegated Regulation (EU) 2022/30. *Off. J. Eur. Union* **2026**, *2026/339*.
14. IEC 62443-4-2; Security for Industrial Automation and Control Systems—Part 4-2: Technical Security Requirements for IACS Components. IEC (International Electrotechnical Commission): Geneva, Switzerland, 2019.
15. ETSI EN 303 645 V3.1.3; CYBER; Cyber Security for Consumer Internet of Things: Baseline Requirements. ETSI (European Telecommunications Standards Institute): Sophia Antipolis, France, 2024.
16. NIST IR 8473; Cybersecurity Framework Profile for Electric Vehicle Extreme Fast Charging Infrastructure. NIST (National Institute of Standards and Technology): Gaithersburg, MD, USA, 2023. [[CrossRef](#)]
17. ISO 15118-20; Road Vehicles—Vehicle to Grid Communication Interface—Part 20: 2nd Generation Network Layer and Application Layer Requirements. ISO (International Organization for Standardization): Geneva, Switzerland, 2022.
18. IEC 61851-1; Electric Vehicle Conductive Charging System—Part 1: General Requirements. IEC (International Electrotechnical Commission): Geneva, Switzerland, 2017.
19. OCPP 2.0.1; Open Charge Point Protocol 2.0.1. OCA (Open Charge Alliance): Arnhem, The Netherlands, 2020.
20. IEC TS 62351-1; Power Systems Management and Associated Information Exchange—Data and Communications Security—Part 1: Communication Network and System Security—Introduction to Security Issues. IEC (International Electrotechnical Commission): Geneva, Switzerland, 2007.
21. Kim, S.; Park, K.-J.; Lu, C. A Survey on Network Security for Cyber-Physical Systems: From Threats to Resilient Design. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 1534–1573. [[CrossRef](#)]
22. Kilic, A. TLS-Handshake for Plug and Charge in Vehicular Communications. *Comput. Netw.* **2024**, *243*, 110281. [[CrossRef](#)]
23. Metere, R.; Neaimeh, M.; Morisset, C.; Maple, C.; Bellekens, X.; Czekster, R.M. Securing the Electric Vehicle Charging Infrastructure. *arXiv* **2021**, arXiv:2105.02905. [[CrossRef](#)]
24. Alcaraz, C.; Cumplido, J.; Triviño, A. OCPP in the Spotlight: Threats and Countermeasures for Electric Vehicle Charging Infrastructures 4.0. *Int. J. Inf. Secur.* **2023**, *22*, 1395–1421. [[CrossRef](#)]
25. Hamdare, S.; Brown, D.J.; Jha, D.N.; Aljaidi, M.; Cao, Y.; Kumar, S.; Kharel, R.; Jugran, M.; Kaiwartya, O. Cyber Defense in OCPP for EV Charging Security Risks. *Int. J. Inf. Secur.* **2025**, *24*, 134. [[CrossRef](#)]
26. Garofalaki, Z.; Kosmanos, D.; Moschogiannis, S.; Kallergis, D.; Douligeris, C. Electric Vehicle Charging: A Survey on the Security Issues and Challenges of the Open Charge Point Protocol (OCPP). *IEEE Commun. Surv. Tutor.* **2022**, *24*, 1504–1533. [[CrossRef](#)]
27. Hamdare, S.; Brown, D.J.; Cao, Y.; Aljaidi, M.; Kumar, S.; Alanazi, R.; Jugran, M.; Vyas, P.; Kaiwartya, O. A Novel Charging Management and Security Framework for the Electric Vehicle (EV) Ecosystem. *World Electr. Veh. J.* **2024**, *15*, 392. [[CrossRef](#)]

28. Biswas, P.; Rashid, A.; Habib, A.K.M.A.; Mahmud, M.; Motakabber, S.M.A.; Hossain, S.; Rokonuzzaman, M.; Molla, A.H.; Harun, Z.; Khan, M.M.H.; et al. Vehicle to Grid: Technology, Charging Station, Power Transmission, Communication Standards, Techno-Economic Analysis, Challenges, and Recommendations. *World Electr. Veh. J.* **2025**, *16*, 142. [\[CrossRef\]](#)
29. IEC TS 62443-1-1; Industrial Communication Networks—Network and System Security—Part 1-1: Terminology, Concepts and Models. IEC (International Electrotechnical Commission): Geneva, Switzerland, 2009.
30. ETSI TS 103 645 V3.1.1; CYBER; Cyber Security for Consumer Internet of Things: Baseline Requirements. ETSI (European Telecommunications Standards Institute): Sophia Antipolis, France, 2024.
31. ETSI TS 103 701 V2.1.1; CYBER; Cyber Security for Consumer Internet of Things: Conformance Assessment of Baseline Requirements. ETSI (European Telecommunications Standards Institute): Sophia Antipolis, France, 2025.
32. NIST CSWP 04162018; Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. NIST (National Institute of Standards and Technology): Gaithersburg, MD, USA, 2018. [\[CrossRef\]](#)
33. Cindrić, I.; Jurčević, M.; Hadjina, T. Mapping of Industrial IoT to IEC 62443 Standards. *Sensors* **2025**, *25*, 728. [\[CrossRef\]](#)
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
35. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
36. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [\[CrossRef\]](#)
37. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223. [\[CrossRef\]](#)
38. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
39. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2312.10997. [\[CrossRef\]](#)
40. 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 the EMNLP, Online, 16–20 November 2020; pp. 6769–6781. [\[CrossRef\]](#)
41. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [\[CrossRef\]](#)
42. NIST AI 100-1; Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST (National Institute of Standards and Technology): Gaithersburg, MD, USA, 2023. [\[CrossRef\]](#)
43. NIST AI 600-1; Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST (National Institute of Standards and Technology): Gaithersburg, MD, USA, 2024. [\[CrossRef\]](#)
44. ISO/IEC 42001; Information Technology—Artificial Intelligence—Management System. ISO (International Organization for Standardization): Geneva, Switzerland, 2023.
45. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [\[CrossRef\]](#)
46. Es, S.; James, J.; Espinosa Anke, L.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta, 17–22 March 2024; pp. 150–158. [\[CrossRef\]](#)
47. Gao, T.; Yen, H.; Yu, J.; Chen, D. Enabling Large Language Models to Generate Text with Citations. In Proceedings of the EMNLP, Singapore, 6–10 December 2023; pp. 6465–6488. [\[CrossRef\]](#)
48. Keskin, F. Memory-Efficient Adaptation and Unanswerable-Aware Supervision for Reliable Private Email Question Answering. *Appl. Sci.* **2026**, *16*, 6645. [\[CrossRef\]](#)
49. Oz, G.; Keskin, F. Perspective Chapter: Operationalizing NIST AI RMF for LLMs—Architecture, RAG and Secure Deployment. In *Artificial Intelligence Ethics, Law, and Policy*; IntechOpen: London, UK, 2026. [\[CrossRef\]](#)
50. Lee, S.-J.; Seo, S.-B.; Bae, Y.-S. A Semi-Automatic Labeling Framework for PCB Defects via Deep Embeddings and Density-Aware Clustering. *Sensors* **2025**, *25*, 6470. [\[CrossRef\]](#)
51. Dalkey, N.; Helmer, O. An Experimental Application of the Delphi Method to the Use of Experts. *Manag. Sci.* **1963**, *9*, 458–467. [\[CrossRef\]](#)
52. Shostack, A. *Threat Modeling: Designing for Security*; Wiley: Indianapolis, IN, USA, 2014.
53. ISO/SAE 21434; Road Vehicles—Cybersecurity Engineering. ISO (International Organization for Standardization): Geneva, Switzerland, 2021.
54. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [\[CrossRef\]](#)

55. Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.V.; Wiest, O.; Zhang, X. Large Language Model Based Multi-Agents: A Survey of Progress and Challenges. In Proceedings of the IJCAI, Jeju, Republic of Korea, 3–9 August 2024; pp. 8048–8057. [\[CrossRef\]](#)
56. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [\[CrossRef\]](#)
57. Fleiss, J.L. Measuring Nominal Scale Agreement Among Many Raters. *Psychol. Bull.* **1971**, *76*, 378–382. [\[CrossRef\]](#)
58. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [\[CrossRef\]](#)
59. Järvelin, K.; Kekäläinen, J. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* **2002**, *20*, 422–446. [\[CrossRef\]](#)
60. McNemar, Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [\[CrossRef\]](#)
61. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, NY, USA, 1993.
62. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [\[CrossRef\]](#)
63. Qwen Team. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115. [\[CrossRef\]](#)
64. Frantar, E.; Ashkboos, S.; Hoefler, T.; Alistarh, D. GPTQ: Accurate Post-Training Quantization for Generative Pre-Trained Transformers. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
65. Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; Han, S. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. *arXiv* **2023**, arXiv:2306.00978. [\[CrossRef\]](#)
66. Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R.J.; Javaheripi, M.; Kauffmann, P.; et al. Phi-4 Technical Report. *arXiv* **2024**, arXiv:2412.08905. [\[CrossRef\]](#)
67. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings through Self-Knowledge Distillation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 2318–2335. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.