photonics

MDPI

Article

LLM-Assisted Standards Conformance Evaluation of Optical Transceivers Through Executable Normative Rules and a Canonical Crosswalk

Haris Hassan [1,2,*], João Pedro [1,3,*] and Luís Cancela [2,3,*]

[1] Nokia, 2790-078 Carnaxide, Portugal
[2] Department of Information Science and Technology, ISCTE—Instituto Universitário de Lisboa, 1649-026 Lisboa, Portugal
[3] Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
* Correspondence: haris.hassan@nokia.com (H.H.); joao.pedro@nokia.com (J.P.); luis.cancela@iscte-iul.pt (L.C.)

Abstract

Qualifying optical transceivers requires interpreting vendor datasheets together with multiple independent standards and multi-source agreements (MSAs), including OIF-400ZR, OpenZR+, OpenROADM, and SFF-8024. Existing retrieval-based approaches can locate relevant clauses, but they do not reliably resolve applicability, identifier namespaces, or executable compliance rules across heterogeneous standards. This paper presents a standards-compilation framework for automated optical transceiver qualification. Normative documents are compiled into versioned executable artifacts that preserve applicability conditions, parameter semantics, units, reference points, and provenance. Proprietary datasheets are converted into structured operating-mode records by a two-pass extractor bound to a typed schema. A canonical crosswalk then maps vendor terms to normative concepts and keeps the original units and reference points. Qualification is performed through deterministic triggering, attribution, and identity-based routing before parameter-level evaluation. The framework is evaluated on commercial coherent optical transceiver datasheets spanning multiple vendors and standards. The two-pass extraction pipeline improves structured identity preservation and reduces the need for reviewer corrections compared with a single-pass baseline. Deterministic routing achieves 99.4% routing accuracy on independently derived reference labels with zero unsafe routing assignments in the evaluated corpus, outperforming retrieval-based baselines. A productivity evaluation further estimates an 85% reduction in expert review time for complex datasheets by shifting manual work from document interpretation to structured evidence review. The proposed framework provides a reproducible and auditable foundation for standards-aware optical transceiver qualification to correct governing records.

Keywords: optical transceivers; standards compliance; interoperability; large language models; retrieval-augmented generation; rule extraction; entity resolution

Check for updates

Received: 2 September 2026
Revised: 24 September 2026
Accepted: 28 September 2026
Published: 30 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Disaggregated optical systems rely on interfaces defined across multiple standards bodies and multi-source agreements (MSAs). A coherent module can simultaneously instantiate several media-interface names, baud rates, optical signal-to-noise ratio (OSNR) figures, and host codes from SFF-8024 [1], as well as marketing claims of conformance to multiple application profiles. These specifications are defined by different authorities.

For example, OIF-400ZR [2] and OpenZR+ [3] define coherent application profiles and operating spaces, OpenROADM [4] defines operational modes together with equipment-level thresholds and penalties, and SFF-8024 [1] assigns the interface code points through which the other documents are identified. Determining whether a transceiver qualifies for a given application therefore requires reasoning across multiple documents rather than consulting one specification.

Nominal membership in a standard family is not evidence of interoperability. For instance, a ten-vendor coherent-pluggable evaluation reported seven failures among 100 endpoint associations, despite the fact that the tested devices supposedly belonged to nominally interoperable 400ZR/OpenZR+ families [5]. Hence, qualification must be performed per parameter, per operating mode, and per revision, with the evidence supporting each decision retained for audit.

Noteworthy, datasheet and MSAs documents resist automation for different reasons. Datasheets advertise several specifications at once but report only a subset of the defined parameters, use vendor terminology, distribute values across tables, prose, and footnotes, and omit conditions implied by a named operating mode profile. Standards define the requirements but express them as revision-specific prose, conditional tables, and cross-references rather than executable constraints. A value, therefore, cannot be judged by matching a parameter name: its applicability depends on revision, application code, operating mode, direction, reference point, measurement condition, and unit.

LLMs are effective at interpreting heterogeneous technical documents and reconciling terminology, and recent applications in the optical networking domain have applied them to orchestration, assistance, and simulation-assisted analysis [6–8]. Direct LLM compliance judgment is nevertheless unsuitable for an auditable process.

LLMs can incorrectly associate values belonging to different operating modes, infer unstated configuration parameters, confuse a vendor conformance claim as a normative bound, or produce a verdict that cannot be reproduced from an explicit rule and evidence trail. Conversely, keyword or embedding retrieval can locate the relevant clause but cannot convert a conditional requirement into an executable check. This provides a motivation for provenance-preserving extraction and reproducible deterministic downstream evaluation for engineering qualifications.

We therefore treat optical transceiver qualification as a standards-compilation problem rather than a document retrieval problem. Normative documents are compiled offline into versioned executable artifacts that preserve applicability conditions, units, reference points, exceptions, and clause-level provenance. Vendor datasheets are converted into typed operating-mode representations through a schema-constrained two-pass extraction process, and a canonical cross-standard taxonomy aligns vendor terminology with normative concepts while preserving measurement semantics. Qualification is then performed through deterministic triggering, attribution, and identity-based routing before parameter-level compliance evaluation.

Our earlier work [9,10] addressed multi-stage datasheet extraction and reported field-level fidelity, scope accuracy, and mode discovery against annotatable fields on 35 datasheets. Using the same decomposition principle (independently prompted subtasks with intermediate validation and restricted context), this paper evaluates what that structured record enables: governing-record routing, estimated expert review time, and compilation compactness. We focus on the normative rules, identity mappings, applicability decisions, and evidence needed to produce a traceable result.

The study makes five contributions:

1. **Versioned standards compilation into executable artifacts**, transforming heterogeneous optical standards and MSAs into versioned rule representations with applicability conditions and provenance (Section 3.3).
2. **A context-sensitive taxonomy and canonical crosswalk** that aligns datasheet fields with standard-specific concepts while preserving units, reference points, and measurement semantics (Section 3.4).
3. A **deterministic routing framework** based on triggering, attribution, and corroboration that identifies the governing standards record before compliance evaluation (Section 3.5).
4. **An evidence-typed verdict model.** Violation, missing evidence, conflict, unverified condition, and out-of-scope are represented separately, and catalog enrichment is kept distinct from vendor-stated evidence (Section 3.6).
5. An **evaluation of four endpoints**: structured extraction quality, governing-record routing, estimated expert effort, and compilation compactness. Parameter-level compliance verdicts are specified in Section 3.6 and are not scored.

2. Background and Related Work

2.1. Machine-Readable Norms and Automated Compliance Checking

Work on machine-readable norms provides the foundations for converting natural-language requirements into computable representations. For instance, LegalRuleML treats rules as first-class objects with provenance, authority, temporal validity, and defeasibility [11]. The RASE methodology separates Requirement, Applicability, Selection, and Exception while preserving contextual information such as inherited headings and table structure [12]. Beach et al. extend these concepts through ontologies, mappings, Semantic Web Rule Language (SWRL) rules, and deterministic procedures for industrial building compliance [13]. Although developed for legal and regulatory reasoning, such structural decomposition is applicable to technical standards, where parameters require evaluation against specific revisions, conditions, explicit exceptions, and cross-referenced clauses. Crucially, this schema reinforces an open-world paradigm, enabling a compliance system to isolate and flag information required by a norm that the source document does not state, instead of silently treating absence as satisfaction [14].

Recent evaluations report that LLMs can identify normative sentences [15], perform compliance classification [14], and extract contract clauses into schema-conforming formal JSON [16]. The transferable lesson is threefold. Semantic models provide stable meaning. Deterministic rules keep decisions reproducible. LLM helps with the expensive conversion from unstructured source material [17]. The remaining step is standards compilation: transforming heterogeneous normative text and tables into validated rules while preserving the relationship.

2.2. Ontology Alignment and Cross-Standard Semantic Mapping

Another distinct challenge is to establish semantic correspondences between heterogeneous schemas and taxonomies. The alignment must keep the meaning and provenance of each source. That alignment is done through a crosswalk. Unlike simple field mapping, a crosswalk provides a reusable semantic bridge that enables interoperability between independently developed knowledge frameworks. It is widely used in metadata standards, knowledge organization systems (KOS), and ontology alignment to integrate information across domains with different vocabularies and structures.

Traditional approaches rely on curated lookup tables, lexical matching, or ontology alignment, which frequently fails on synonymous labels, polysemous terms, and independently evolving vocabularies. Recent advancements mitigate these challenges through

LLM-assisted taxonomy alignment. These frameworks infer semantic equivalence by exploiting hierarchical context and projecting disparate documents onto a shared set of semantic aspects prior to evaluation [18,19].

Optical standards present a more constrained but stricter instance of this problem: the same physical quantity may appear under different names or measurement reference points, while identical numeric identifiers may belong to different namespaces. This motivates a canonical crosswalking layer that aligns standards through vocabulary-neutral engineering concepts while maintaining traceability to each authoritative specification. Such a representation supports interoperable retrieval, revision-aware mapping, and deterministic standards qualification without collapsing multiple standards into a single ontology.

2.3. LLM and AI in Optical-Network Automation

Optical-network research has demonstrated that LLMs can assist configuration, optimization, and domain reasoning. Zaid et al. present an on-premises multi-agent assistant grounded in configuration, inventory, telemetry, datasheets, and YANG models, with explicit audit trails and safe-action previews [6]. Research shows that structured physics- and system-aware context improves a compact on-premises model in an optical optimization task [7]. Niu et al. introduce a vertically pretrained optical foundation model and a task-specific benchmark, illustrating the value of domain-specific data and evaluation [8].

Digital-twin work emphasizes independent validation: Shen et al. place simulation inside a traceable intent-to-action loop [20]. Delezoide et al. propose physics and software validation around an LLM-driven workflow and stress that twin uncertainty must be quantified [21]. Arunachalam et al. identify physics-based guardrails as necessary because generative models do not intrinsically enforce optical constraints [22]. Other systems combine LLMs with retrieval, APIs, digital twins, grammars, planning languages, or solvers [23,24].

These systems reason over formalized knowledge and assume that the standards, planning domains, or compliance constraints have already been formalized. To the best of our knowledge, the upstream problem—constructing and validating that formal representation from heterogeneous normative sources and doing so in a way that can be audited and re-executed against a specific revision—has not been comprehensively addressed in the literature.

2.4. The Identified Gap

Optical AI work targets network operation, planning, optimization, and general document assistance; automated-compliance work targets legal, financial, and building regulation. Neither provides a system that compiles heterogeneous optical standards and MSA tables into executable rules, resolves identifiers across the SFF-8024, OIF, OpenZR+, and OpenROADM namespaces, routes a vendor transceiver profile to the correct rule subset, and produces parameter-level verdicts with source-to-rule-to-datasheet provenance and explicit missing-evidence semantics. Table 1 positions this work.

Table 1. Positioning against the adjacent literature.

Capability	Optical Network Standards	Automatic Compliance Checking	Taxonomy Alignment	This Work
Normative source compiled offline into artifacts	–	✔	–	✔
Requirements executable with units and reference points	–	partial	–	✔
Applicability as multi-axis configuration coordinates	–	–	–	✔

Table 1. *Cont.*

Capability	Optical Network Standards	Automatic Compliance Checking	Taxonomy Alignment	This Work
Identity resolution across independent namespaces	–	partial	✔	✔
Deterministic parameter-level verdict	–	✔	–	✔
Clause-level provenance and revision identity	partial	✔	partial	✔
Explicit abstention/missing-evidence semantics	–	partial	–	✔

3. Problem Formulation and Proposed Framework

Optical transceiver qualification is formulated as a structured evidence-to-compliance problem over two heterogeneous document classes: vendor datasheets, which describe a device's operating modes, and normative specifications, defining the requirements governing interoperable operation. The challenge is not extracting parameter values alone but determining which normative requirements apply before those values are compared.

The framework therefore separates qualification into three deterministic stages:

1. Evidence extraction—recover operating modes from the transceiver datasheet.
2. Standards routing—identify governing standards records for each operating mode.
3. Compliance evaluation—compare extracted parameters against the applicable normative requirements.

This separation prevents document interpretation, applicability resolution, and numerical validation from being conflated into a single LLM judgment. Figure 1 provides a schematic representation of the framework.

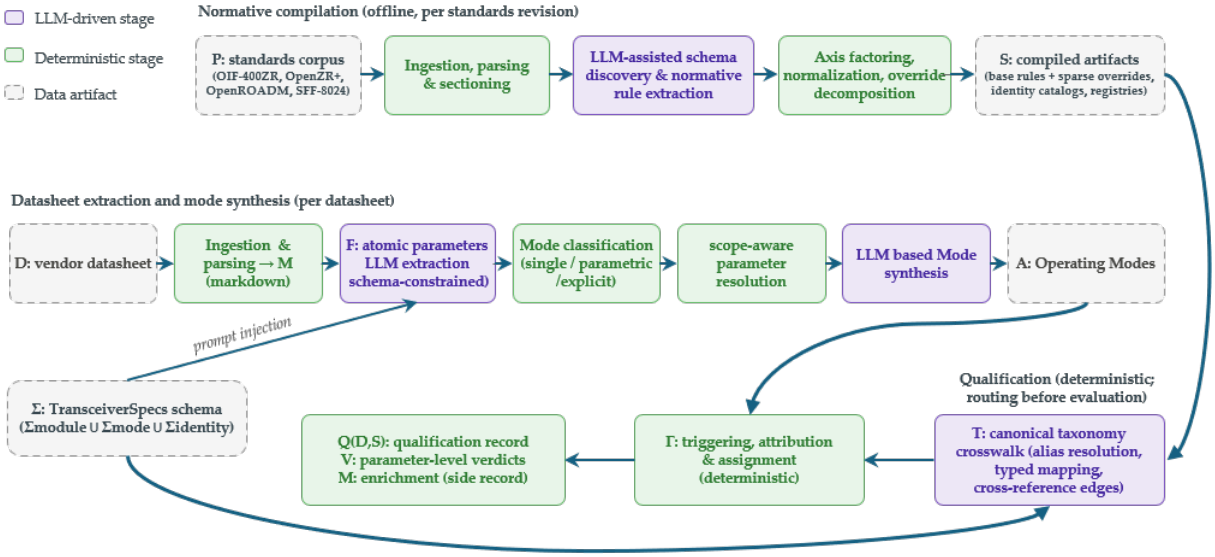


Figure 1. Proposed framework. Purple stages use an LLM; green stages are deterministic.

Let D denote a vendor datasheet, P the source normative documents, S the compiled standards corpus, Σ the transceiver evidence schema, and T the canonical crosswalk. Evidence extraction maps D to atomic facts F and datasheet operating modes A. Applicability resolution maps A, S, and T to the governing assignments Γ. Deterministic evaluation then produces the qualification record Q. Equations (1)–(5) summarize this sequence.

$$D \xrightarrow[\Sigma]{\textit{evidence extraction and mode synthesis}} (F, A) \tag{1}$$

$$P = \{p_1, \ldots, p_n\} \overset{\textit{normative compilation}}{\rightarrow} S = \{s_1, \ldots, s_n\} \tag{2}$$

$$\{s_1, \ldots, s_n\}; \Sigma \overset{\textit{semantic resolution \& deterministic processing}}{\rightarrow} T \tag{3}$$

$$(A,\ S,\ T) \overset{\textit{triggering, attribution \& assignment}}{\rightarrow} \Gamma \tag{4}$$

$$(A, \mathrm{S}, \mathrm{T}, \Gamma) \overset{\textit{deterministic evaluation}}{\rightarrow} Q \tag{5}$$

The framework comprises six components, described in Sections 3.1–3.6: the evidence schema, datasheet extraction and mode synthesis, normative compilation, canonical crosswalking, applicability resolution, and deterministic compliance evaluation.

3.1. Transceiver Specifications Schema Using Pydantic

The transceiver specification schema, denoted by Σ, defines the canonical evidence model used throughout the qualification framework. Implemented as a Pydantic model V2, it serves as the single source of truth for evidence extraction, validation, storage, and downstream deterministic processing. Every extracted parameter is represented once in Σ, with an explicit field name, data type, optionality, unit, and natural-language description. Hence, both the LLM extraction prompt and the validation pipeline operate against a centralized, formally defined interface rather than separate parameter lists.

The schema fulfills two complementary functions. First, it provides the extraction contract presented to the LLM. Instead of embedding a manually curated list of parameters into the prompt, the framework injects a compact JSON schema generated directly from the Pydantic model using model_json_schema(). Field names, types, descriptions, and units therefore become the instructions seen by the model. Second, it provides the acceptance contract for extracted evidence. The LLM response is parsed as structured JSON, minor scalar-list inconsistencies are coerced where appropriate, and the result is validated against the same Pydantic class before entering the qualification pipeline. Any subsequent human edits are subjected to the identical validator, ensuring that manually reviewed evidence satisfies the same structural constraints as automatically extracted evidence.

The schema partitions fields by scope and separately tags those used for identity resolution, as formalized in equation

$$\Sigma = \Sigma_{module} \cup \Sigma_{mode} \tag{6}$$

Module-level fields describe properties that characterize the physical transceiver independently of operating configuration, including vendor, model identifier, form factor, connector type, supported fiber type, operating temperature range, and other documentary attributes that remain constant across all operating modes.

Mode-level fields represent parameters whose values may differ between supported operating configurations. Examples include bit rate, baud rate, modulation format, FEC type, channel spacing, transmission reach, optical signal-to-noise ratio (OSNR), and host or media interface identifiers. These fields may be single-valued, ranges, or explicitly multi-valued within a single datasheet.

Identity-bearing fields are those capable of selecting or qualifying a standard-defined operating mode. They include application codes, standards claims, host interface names and code points, media interface names and code points, and other identifiers used during applicability resolution. Marking these fields explicitly allows later stages to distinguish between descriptive operating parameters and parameters that establish normative identity.

In addition to typed technical parameters, Σ contains three documentary fields that preserve information that cannot always be represented by structured scalar values:

- **Notes**: Captures observations or specifications that do not fit an existing typed field.
- **Suggestions**: Records potential schema limitations or proposed extensions identified during extraction.
- **provenance_datasheet_sections**: Records the datasheet section headings containing the extracted evidence.

These documentary fields act as controlled escape valves rather than allowing unsupported completion. When a specification cannot be represented by the current schema—for example, a storage-temperature specification when only operating temperature is modeled—the observation is preserved explicitly instead of being forced into an incorrect field or discarded. Likewise, repeated suggestions across multiple datasheets provide evidence for controlled schema evolution rather than silent extraction failures.

The schema treats every extracted parameter as an atomic, source-grounded fact rather than an inferred property. Numeric values are represented using their documentary units, categorical parameters retain their declared vocabulary, and optional fields distinguish between an absent value and an extracted null. This representation intentionally separates evidence recovery from semantic interpretation: extraction records what the datasheet states, while later stages determine applicability and compliance.

The schema also distinguishes between scalar and collection-valued parameters. Fields representing multiple supported values—such as supported modulation formats or multiple host interfaces—are modeled as lists rather than flattened strings, allowing downstream algorithms to reason over individual values deterministically.

Because every downstream component references fields through their canonical schema paths, Σprovides a stable interface between evidence extraction and deterministic qualification. The canonical crosswalk associates canonical concepts with these schema field paths, while mode synthesis partitions extracted values according to their schema category without introducing alternative representations. This design makes schema evolution explicit and auditable. Adding, renaming, or removing a field changes the generated extraction schema and automatically exposes unresolved downstream mappings during validation, preventing silent incompatibilities between extraction, routing, and compliance evaluation. The resulting evidence model provides a consistent typed representation of vendor datasheets that is independent of any individual standard while remaining sufficiently structured for deterministic processing throughout the remainder of the framework.

3.2. Datasheet Parsing, Extraction, and Mode Synthesis

Let D denote the vendor datasheet PDF. Rather than querying the PDF directly, the document is converted into an editable Markdown representation:

$$\left(D \overset{Parser}{\rightarrow} M\right) \tag{7}$$

Markdown preserving enough document structure for both LLM inference and operator review. A human revision can be done to correct any incorrectly parsed parts of PDF before evidence extraction begins. The first extraction pass recovers typed evidence using the schema defined in Section 3.1. The extraction prompt receives only the Markdown document and the JSON schema generated from Σ

$$Prompt(M;\Sigma) \overset{LLM}{\rightarrow} F \tag{8}$$

LLM returns F under an explicit contract: copy numeric values as printed, and do not invent, convert, or round; emit null when a value is absent; treat each field as an independent atom and do not infer one from another; and for list fields collect every distinct value occurring in M. The output is a collection of atomic facts F whose provenance

remains tied to the original datasheet. Alongside typed parameters, the extractor also returns documentary notes, suggested schema extensions, and the section headings from which evidence was obtained. These provenance headings are later used to restrict the context provided during mode synthesis to only the relevant portions of the datasheet.

After validation, parameters are classified according to the schema categories. Module-level parameters remain constant across the device, whereas mode-level parameters are analyzed to determine the configuration structure of the transceiver. The framework distinguishes three classes of transceivers according to the multiplicity of Σ_{mode} values:

- **single-configuration**, in which every mode-level field contains one value (e.g., a device specified only for a fixed 10G SR configuration);
- **explicit multi-configuration**, in which multiple finite discrete values describe distinct operating modes (e.g., separate 100, 200, and 400 Gb/s configurations).
- **parametric**, in which mode-level fields contain ranges, increments, or tunable parameter values (e.g., Nokia CHM7 with tunable dimensions—capacity, baud rate, roll-off, wavelength, and modulation—and configurable increments that define a configurable operating space);

For single-configuration modules, operating mode reconstruction is trivial because every configuration-dependent parameter already describes the unique supported mode:

$$A_{single} = F(\Sigma_{module}) \cup F(\Sigma_{mode}) \tag{9}$$

For modules supporting multiple configurations, extracted facts are separated into module-scoped and configuration-scoped evidence.

Module-scoped evidence consists of all module-level fields together with mode-level parameters that contain only a single value and therefore apply uniformly across every configuration. Configuration-scoped evidence contains only the genuinely multi-valued mode parameters whose relationships remain unresolved after extraction. Formally,

$$F_{module} = F(\Sigma_{module}) \cup F_{single}(\Sigma_{mode}) \tag{10}$$

$$F_{mode} = F_{multi}(\Sigma_{mode}) \tag{11}$$

This partition prevents constant parameters from being duplicated across operating modes while preserving unresolved configuration choices for the synthesis stage. For parametric modules, the extracted ranges already describe the supported operating space, so the framework reports the parameterized configuration without attempting to enumerate discrete modes.

Explicit multi-configuration modules require reconstruction of complete operating modes because datasheets commonly distribute related parameters across separate tables or specification sections. A second LLM pass receives only the unresolved mode-scoped parameters F_{mode} together with the provenance-identified sections of the Markdown M_{prov}:

$$Prompt(F_{mode}, M_{prov}) \overset{LLM}{\rightarrow} A_{explicit} \tag{12}$$

The synthesis objective differs fundamentally from extraction. Rather than discovering new parameter values, the model resolves which extracted values co-occur within the same documented operating configuration. A synthesized mode is accepted only when the datasheet provides explicit evidence linking the participating parameters. If multiple values cannot be associated unambiguously, for example, when two FEC values appear without indicating which modulation format each accompanies—the parameter remains unresolved instead of being combined speculatively. The synthesis pass reconstructs

only documented configurations rather than generating plausible combinations. Each synthesized operating mode is then combined with the module-scoped evidence:

$$A_{explicit} = F_{module} \cup A_{explicit}(F_{mode}) \tag{13}$$

The two-pass design separates evidence recovery from configuration reconstruction. Extraction guarantees that every value originates directly from the datasheet, whereas synthesis determines how those independently extracted values form valid operating modes. This distinction is necessary because syntactically valid structured output does not guarantee semantic consistency. For example, if extraction identifies two distinct FEC values, treating them as interchangeable during extraction could fabricate operating modes not described by the vendor. By introducing an intermediate deterministic partition, unresolved ambiguity is preserved until sufficient documentary evidence exists to resolve it.

The extraction stage produces three versioned-intermediate artifacts that decouple document processing from qualification:

- specs.json: Validated atomic facts F extracted under the schema.
- mode_atoms.json: Partition of module-scoped and mode-scoped evidence.
- modes.json: Synthesized operating modes A used by downstream qualification.

Subsequent stages consume these structured artifacts directly and never re-parse the original datasheet. This makes qualification deterministic with respect to the extracted evidence while preserving complete traceability back to the reviewed Markdown representation and the original vendor document.

3.3. Normative Compilation

Each normative source p is compiled into a versioned executable structured artifact s. A new revision of p therefore requires recompilation before the revised requirements can be used. This separation provides deterministic execution, version control, provenance, and reproducibility. In this work

$$S = \{\text{OIF} - 400\text{ZR},\ \text{OpenZR+},\ \text{OpenROADM},\ \text{SFF} - 8024\} \tag{14}$$

The four sources represent different kinds of normative knowledge and therefore produce different executable artifact types rather than a single uniform representation as described in Table 2.

Table 2. Source-specific executable artifacts.

Normative Source	Compiled Artifact	Primary Purpose
OIF-400ZR	Axis-factored rule store	Configuration-dependent normative requirements with sparse overrides.
OpenZR+	Axis-factored rule store	Multi-dimensional configuration rules spanning payload, modulation, power profile, grid spacing, and add/drop variants.
OpenROADM	Operational-mode catalog	Identity-scoped operational modes whose limits are attached to catalog records.
SFF-8024	Identity registry	Host and media interface identifiers, code points, and naming registries used for identity resolution.

Compilation procedure, per source kind. Compilation follows different procedures depending on the source type, but every artifact passes through a deterministic normalization stage before entering the versioned compiled standards corpus. For SFF-8024, registry tables are extracted and normalized while preserving assigned names, hexadecimal code

points, notes, and metadata. The resulting artifact functions as a deterministic lookup table for host and media interface identities. For OpenROADM, the published operational-mode catalog is imported directly into the executable representation. Each catalog row represents a complete operating identity rather than an independent parameter list, allowing later stages to resolve a datasheet mode to a single catalog record before evaluating associated limits. For OIF-400ZR and OpenZR+, normative PDFs require semantic extraction because requirements are distributed across prose, tables, notes, and conditional exceptions. These documents are transformed into an intermediate structured representation using provenance-preserving, section-scoped, multi-pass LLM extraction.

Each extracted normative rule is represented using the four RASE roles—Requirement, Applicability, Selection, and Exception—which separate the value being constrained from the conditions under which it applies. Selection denotes an alternative operational branch or configuration scope rather than a recommendation.

Beyond the RASE roles, every rule records the structured information required for deterministic evaluation, including:

- normative modality (shall, shall not, should, may);
- subject and canonical parameter candidate;
- comparison operator and required value;
- unit and reference bandwidth where applicable;
- physical reference point;
- operational-mode constraints and applicability conditions;
- exceptions and footnotes;
- document cross-references; and
- source span identifying the originating clause, table, or paragraph.

This intermediate representation preserves the distinction between normative content and documentary provenance. Every executable rule remains traceable to its originating document location.

Intermediate representation is then subjected to **deterministic validation** and reconciliation before entering the versioned source of truth: JSON-schema validation; quantity and unit compatibility; operator/value compatibility; physical reference-point compatibility; source-span existence; cross-reference resolution; and detection of duplicate, inconsistent, or contradictory records. Low-confidence mappings require human approval. The LLM is thus constrained to *proposing* evidence-backed rule structure over bounded source material; it does not decide compliance.

Axis factoring. A central challenge in compiling optical standards is that many requirements depend on combinations of operational configuration dimensions. Representing every configuration independently produces large, highly redundant rule tables. A dense representation defines an independent rule for every coordinate tuple. The framework instead represents requirements using explicit configuration axes together with sparse conditional overrides. The relevant axes differ by specification. OIF-400ZR varies primarily by application code, while OpenZR+ introduces multiple simultaneous axes, including payload rate, modulation format, symbol rate, power profile, channel spacing, add/drop configuration, and frequency grid spacing.

Instead of encoding requirements according to table position, the compiled artifact records named applicability conditions. For example, an override is represented as "application code equals 0×03" $\times$ rather than "third table column." This converts document layout into executable conditions.

Base rules and sparse overrides. Each canonical parameter is represented as

$$R_p = \left(R_p^{\{base\}}, O_p\right),\ O_p = \{(c_i, r_i)\} \tag{15}$$

where $R_p{}^{base}$ is the default requirement and each condition c_i ranges over one or more named axes. At evaluation time the base rule is selected first and is replaced or refined by an override whose condition matches the resolved configuration:

$$R_p(X) = R_p^{\{base\}} \oplus \left\{ O_{\{p,i\}} \mid c_i(X) = true \right\} \tag{16}$$

where $\oplus$ denotes deterministic application of matching overrides. Only overrides whose conditions are satisfied modify the inherited base requirement. If multiple matching overrides impose incompatible requirements, the parameter is marked indeterminate rather than resolved heuristically. For example, if a chromatic-dispersion requirement is 2400 ps/nm for application codes 0×01 and 0×03 but 1200 ps/nm for 0×02, the compiled artifact retains 2400 ps/nm as the base requirement and encodes only the exceptional configuration as a scoped override.

Inheritance. Many normative parameters do not vary across operational configurations. Instead of repeating identical values for every configuration tuple, the artifact represents inheritance explicitly. A parameter without a matching override inherits its base rule throughout the applicable configuration space. Parameters such as post-FEC BER, specified fiber type, optical return loss, or polarization rotation speed therefore appear only once unless the specification defines an alternative configuration-specific requirement. This makes inheritance an explicit semantic property of the artifact rather than an implicit consequence of missing table entries. The evaluator can distinguish between "inherits the base requirement" and "no requirement exists."

Identity catalogs. Some standards define operational identities whose semantics depend on combinations of related protocol and physical attributes rather than individual scalar parameters. The compiled standards corpus therefore includes row-level identity catalogs alongside parameter rules. Each catalog record represents a complete operational identity that may simultaneously specify application code, host interface, media interface, modulation format, lane mapping, FEC, symbol rate, payload rate, and other identifying characteristics. Treating these rows as immutable identity objects allows applicability resolution to match an extracted operating mode against a single catalog profile before retrieving its associated normative limits. This is particularly important for OpenROADM operational modes and identity-bearing OpenZR+ or SFF-8024 interface definitions.

Every qualification outcome is explicitly tied to the versions of the standards used to produce it. Historical modules can therefore be evaluated against previous standard revisions, while newer revisions become available only through recompilation of the corresponding source document. This versioned compilation stage separates normative interpretation from compliance evaluation and guarantees that downstream qualification operates on deterministic, provenance-preserving executable artifacts rather than directly interpreting heterogeneous specification documents.

3.4. Canonical Taxonomy Crosswalk

The canonical crosswalk, denoted by T, is the semantic join layer between extracted datasheet evidence and compiled normative artifacts. Vendor datasheets and standards describe the same engineering concepts using different terminology, units, identifiers, and document structures. The crosswalk resolves these heterogeneous representations into a vocabulary-neutral set of canonical concepts without modifying the underlying evidence or normative rules. Rather than acting as a synonym table, T preserves the semantic information required for deterministic routing and comparison. Each canonical concept records a preferred label, quantity kind, canonical SI unit, dimensional category, comparison convention, optional reference point, and provenance links to the corresponding vendor and standard representations.

A **canonical concept** represents one engineering quantity independently of how individual sources describe it. Each concept maintains mappings to one or more datasheet fields and one or more standard-specific parameters while preserving source identity and document provenance. This representation separates semantic identity from documentary vocabulary. For example, vendor-specific parameter names, OIF terminology, OpenZR+ field names, and OpenROADM catalog attributes may all refer to the same canonical concept while remaining distinguishable in their original sources.

Construction: The crosswalk is constructed offline from the compiled standards corpus rather than during qualification. Candidate parameter names, aliases, definitions, units, reference points, and source paths are collected from each supported standard. Semantic embeddings identify groups of potentially equivalent parameters, after which an LLM evaluates each candidate using its surrounding definition, terminology, units, and measurement context. Candidate mappings may be accepted, rejected, or left unresolved.

Accepted mappings are then normalized deterministically by assigning canonical units, quantity kinds, and comparison conventions. Parameters that cannot be mapped confidently are retained as unresolved concepts rather than being forced into an existing category.

Deterministic routing: At qualification time, routing is performed through exact schema field paths rather than semantic similarity. On the datasheet side, each field in Σ maps to one or more canonical concepts. On the standards side, each canonical concept maintains reverse mappings to the corresponding executable rule or catalog location. The crosswalk therefore realizes the deterministic mapping.

$$\textit{Datasheet Field} \rightarrow \textit{Canonical Concept} \rightarrow \textit{Standard Parameter} \rightarrow \textit{Applicable Rule}$$

A field extracted from a datasheet but absent from T is classified as unrouted, preserving the evidence for review instead of silently discarding it. Multiple mappings are retained only when a field legitimately corresponds to multiple canonical concepts.

Semantic alignment versus numerical comparability: Semantic correspondence does not imply that two values can be compared for compliance. The crosswalk establishes that two source representations refer to the same engineering concept, while numerical comparability is evaluated separately using retained metadata such as units, reference bandwidths, physical reference points, and operating conditions. For example, in-band OSNR is represented by corresponding parameters in OIF-400ZR, OpenZR+, and OpenROADM, but differences in reference bandwidth or measurement definition may prevent direct numerical comparison. By preserving this metadata within the crosswalk, the framework prevents semantic similarity from being mistaken for compliance evidence.

The canonical crosswalk therefore provides a stable semantic interface between extracted evidence and executable standards while preserving the contextual information required for deterministic applicability resolution and compliance evaluation.

3.5. Applicability Resolution: Triggering, Attribution, and Assignment

Compliance evaluation begins by determining which standards and operational modes govern each extracted transceiver mode. This applicability resolution stage operates before numerical evaluation and consists of three deterministic steps: standard triggering, mode attribution, and standard–mode assignment. Separating these steps prevents parameter similarity from being interpreted as evidence of conformance and provides an auditable path from datasheet evidence to the normative rules subsequently evaluated.

Candidate identity profile: For each synthesized operating mode $a \in A$, the framework constructs a candidate identity profile

$$\pi(a) = \langle c_a, c_h, c_m, l_h, l_m, r_h, r_m, \mu, FEC, G \rangle \tag{17}$$

where c_a, c_h, and c_m denote application, host, and media identifiers; l denotes lane mappings; r denotes host and media rates; μ denotes modulation format; FEC denotes forward-error correction; and G denotes grid or channel information. Only fields present in the extracted evidence contribute to the identity profile.

Standard triggering: The compliance engine evaluates each candidate standard against the identity profile and returns

$$CE(a, s) = (v, t, E) \tag{18}$$

where $v \in$ {applicable, not applicable, indeterminate} is the applicability verdict of 5 standards artifacts over datasheet operating mode, t is the trigger type, and E contains the supporting evidence. Three trigger types are recognized:

- **Explicit:** the datasheet directly claims a standard or provides an identity-bearing code point defined in $\Sigma_{identity}$.
- **Inherited:** applicability follows from a normative dependency on another already-triggered standard.
- **Implicit:** no explicit claim exists, but the extracted identity profile uniquely matches a standard-defined operational mode.

An applicable verdict requires positive supporting evidence, while a not applicable verdict is returned only when available evidence explicitly excludes the standard. Missing or incomplete evidence produces indeterminate results, ensuring that absence of evidence is not interpreted as evidence of non-applicability.

Mode attribution: Triggering establishes that a standard may govern a mode; attribution determines which operational mode within that standard corresponds to the extracted configuration. This step is required because multiple catalog modes may share non-identifying characteristics, such as the same host electrical interface or channel spacing.

Let V_a denote the verified identity-bearing values extracted for mode a, and let $I_{s,i}$ denote the identity fields of catalog entry i in standard s. A catalog mode is considered a candidate only when identity-bearing evidence overlaps:

$$V_a^{ID} \cap I_{s,i}^{ID} \neq \varnothing \tag{19}$$

Identity-bearing fields include media interface names and code points, application codes, standards codes, and other unique identifiers. Parameters shared across multiple catalog modes, such as a common host interface or symbol rate, may corroborate a match but cannot independently identify a mode.

Identity verification and assignment: Candidate identities are verified using deterministic comparison of names, code points, and supporting operational attributes. Code-point representations are normalized during comparison so equivalent hexadecimal or decimal representations match without altering the recorded evidence. Limited numeric tolerance is permitted where datasheets round values such as symbol rates.

A numeric identifier alone is insufficient evidence if multiple registries reuse the same value. e.g., an OIF application code, an SFF-8024 host code, and an SFF-8024 media code that share a numeral are different keys. A match is admitted only when the extracted field path and a second cue (interface name, explicit standard claim, or already-triggered parent standard) place the numeral in one namespace. Code-point spelling is normalized

(hex/decimal); name comparison folds case and punctuation but retains "+", so 400ZR+ and 400ZR remain distinct. Worked collision: application code 0×01 is a valid OIF-400ZR key and also SFF-8024 media row 0×01 (10GBASE-SW). Identifier-only matches therefore require corroborating evidence, such as a matching interface name or an explicit standards claim.

Once a unique catalog profile is verified, the framework records the assignment

$$(a, s_i, A_{s_i,j})$$

where a is the extracted operating mode, s_i is the applicable standard, and $A_{s_i,j}$ is the resolved standard-defined operational mode or catalog profile. The assignment also stores the trigger type and supporting evidence used to establish the match.

Applicability output: Applicability resolution produces the set of governing standards and operational-mode assignments that drive deterministic compliance evaluation. Parameters retrieved from identity catalogs may enrich the evaluation context, but catalog-derived information is stored separately from datasheet evidence so the provenance of every compliance input remains explicit. Only after a standard–mode assignment has been established are the corresponding normative rules selected for numerical evaluation.

3.6. Verdict Model and Findings

After applicability resolution, the framework evaluates extracted evidence against the normative requirements associated with each assigned standard–mode pair. Compliance evaluation is deterministic: the applicable rules are selected from the compiled standards corpus, routed through the canonical crosswalk, and compared with the corresponding datasheet evidence without further LLM reasoning. For a vendor datasheet D, compiled standards S, and taxonomy T, the qualification record is

$$Q(D, S, T) = \{(p, f, r, v, e)\} \tag{20}$$

where p is the canonical parameter, f is the extracted datasheet evidence, r is the applicable normative rule, v is the evaluation verdict, and e records the supporting provenance.

Parameter-level verdicts: Each applicable rule is evaluated independently after unit normalization and compatibility checks. The framework produces one of four parameter-level verdicts.

- **Conformant:** Comparable evidence exists and satisfies the applicable requirement.
- **Non-conformant:** Comparable evidence exists and violates the applicable requirement.
- **Indeterminate:** The requirement applies, but evaluation cannot be completed because evidence is missing, unresolved, or not directly comparable.
- **Not specified:** The resolved standard contains no applicable scalar requirement for the selected profile, or the artifact is descriptive rather than directly checkable.

Indeterminate evaluations arise from missing datasheet values, unresolved multi-valued parameters, unit or reference-point incompatibilities, blocked selector axes, or conflicting applicable overrides.

Evidence and provenance semantics: The framework distinguishes missing evidence from technical non-compliance. A missing datasheet value does not automatically produce a failure; if the corresponding rule applies but the required evidence is absent, the verdict remains indeterminate unless the standard explicitly requires that the value be declared. Similarly, enrichment does not replace missing datasheet evidence. Information retrieved from standards catalogs may provide additional context for evaluation, but it is retained as catalog-derived evidence and never converted into an observed vendor specification.

Provenance therefore distinguishes values stated in the datasheet from values obtained through normative enrichment. Explicit standards claims are also treated as descriptive evidence. They may contribute to applicability resolution and identity assignment, but they do not establish conformity with the requirements of the claimed standard.

Qualification findings: The qualification record is accompanied by human-readable findings that explain the outcome of deterministic evaluation. Findings are grouped into five categories:

- **Compliance verdicts** report successfully evaluated requirements together with their normative references.
- **Threshold violations**, identifying comparable values that violate applicable limits.
- **Missing information**, identifying evidence required to complete an applicable evaluation.
- **Conflict detection**, identifying inconsistent or mutually incompatible identity evidence that blocks assignment or evaluation.
- **Technology classification**, providing descriptive interpretation of the extracted configuration (for example, identifying consistency with a legacy coherent implementation rather than a 400ZR profile).

This separation ensures that technical violations, missing evidence, identity conflicts, and descriptive observations remain distinct outcomes rather than being collapsed into a single pass/fail result. The resulting qualification record is therefore an auditable collection of parameter-level compliance decisions together with the evidence and provenance supporting each decision.

4. Experimental Methodology

The experiments address four questions:

1. **RQ1—input quality:** Does two-pass extraction recover an operating mode-aware record with preserved typed identity from datasheets more often than a single-pass call?
2. **RQ2—expert effort:** How much expert time does the automated extraction and mode synthesis remove?
3. **RQ3—routing:** Does canonical crosswalk and deterministic routing identify the governing record more reliably than retrieval-based selection?
4. **RQ4—compilation fidelity:** What are the cross-standard mapping structure, provenance traceability, and factoring reduction of the executable standards corpus?

RQ3 evaluates routing rather than parameter conformance because every downstream parameter verdict depends on the selected governing record. RQ1 supplies the identity evidence consumed by that routing step.

4.1. Corpus

Datasheet corpus (RQ1, RQ2). As shown in Table 3, the evaluation used 50 optical-transceiver datasheets from seven vendors covering 111 operating modes.

Routing item set (RQ3). Each operating mode was evaluated against five routing targets, as shown in Table 4, producing ($111 \times 5 = 555$) mode–target evaluation items. Host and media interfaces are evaluated independently because a transceiver may specify them separately.

Table 3. Datasheet corpus used for extraction, effort analysis, and routing.

Quantity	Value
datasheets	50
vendors	7
modes	111

Table 4. Routing targets and candidate records in the compiled normative artifacts.

Target	Answer Vocabulary	Records
OIF-400ZR	application code (0 × 01–0 × 03)	3
OpenZR+	identity name, e.g., ZR400-OFEC-16QAM	14
OpenROADM	operational-mode id, e.g., OR-W-400G-oFEC-118Gbd	22
SFF-8024/host	host electrical interface identifier	256
SFF-8024/media	media interface identifier	513

Table 5 summarizes governed datasheet operating modes and positive assignments by target. 62 modes governed by at least one target produce 86 positive mode–target assignments because a mode may legitimately map to multiple targets.

Table 5. Distribution of datasheet operating modes and positive routing assignments by target.

Target	Modes Governed	Positive Routing Assignments
OIF-400ZR	10	10
OpenZR+	15	15
OpenROADM	3	3
SFF-8024/host	48	48
SFF-8024/media	10	10
At least one target	62	86
No applicable target	49	—

Retrieval corpora. The prose corpus contains 698 chunks from 326 pages: 244 from OIF-400ZR, 183 from OpenZR+, 182 from OpenROADM, and 89 from SFF-8024. The record corpus contains 900 chunks, of which 807 identify a normative record: 768 from SFF-8024, 76 from OpenZR+, 34 from OIF-400ZR, and 22 from OpenROADM.

4.2. Reference Standard

Routing reference labels were derived independently from the compiled normative registries rather than from the canonical crosswalk, preventing the evaluation labels from depending on the resolver under test. Matching uses deliberately conservative rules: element-wise name equality, exact identifier equality, and an explicit standard claim before an identifier is trusted. Reserved, vendor-specific, undefined, and unallocated SFF-8024 codes could not establish a governing record.

Each item receives one of three labels: a specific governing record; NONE, when the available evidence establishes that nothing in that registry governs the item; or REVIEW, when a human could propose a record but the documentary identity evidence does not

establish it uniquely. The 83 REVIEW items are withheld from scoring. The 83 REVIEW items comprised 46 OpenZR+ claims without an explicit identifier, 13 similarly inferable OpenROADM claims, 19 qualified names matching more than one registry row, and five claims with a missing or non-unique application code. REVIEW items were withheld from scoring rather than resolved by inference. Distribution is shown in Table 6.

Table 6. Routing reference labels.

Category	Items
Total routing items	555
Scoreable	472
governed by a record	86
governed by nothing (NONE)	386
Review	83

During scored evaluation, a system returned either a specific record or an abstention. If the implementation token is NONE, identify it as predicted NONE in the output.

4.3. Systems, Baselines, and Ablations

Table 7 lists the configuration of the proposed system, three retrieval baselines, and ablations of the proposed system.

Table 7. Routing systems: proposed method P, retrieval baselines B1–B3, and ablations A1, A2, and P′.

ID	Configuration	Isolates
P—crosswalk	compiled standards + canonical crosswalk + corroboration + deterministic routing	the proposed system
P′—no corroboration	P with the corroboration predicate disabled, so a bare identifier suffices.	the contribution of one rule
B1—prose retrieval	Dense retrieval over the published PDFs, page-window chunks, then model adjudication.	source representation
B2—record retrieval	Dense retrieval over the compiled standards, one chunk per record, then model adjudication	selection method, corpus held constant
B3—hybrid retrieval	Hybrid BM25 + dense retrieval, cross-encoder reranker, schema-constrained decoding	retrieval method
A1—no crosswalk	P with crosswalk's name join replaced by literal string equality, keeping deterministic selection.	the name join
A2—model adjudication	P with the crosswalk but LLM adjudication instead of deterministic routing	contribution of determinism

The retrieval arms used the same six-excerpt budget, answer vocabulary, and decoding configuration where applicable. B1 used dense retrieval over published-PDF prose, whereas B2 used dense retrieval over compiled record chunks; their comparison therefore changes the representation and retrieval unit. B2 and B3 compare complete retrieval stacks—dense retrieval versus hybrid retrieval with fusion and reranking—over the compiled record corpus. P and B2 share the compiled input files and evaluation task, but their comparison is a composite system comparison that changes crosswalk use, corroboration, retrieval, and model adjudication.

4.4. Evaluation Metrics

The evaluation measures four research questions corresponding to the major stages of the framework. Each stage is evaluated independently using metrics aligned with its objective.

RQ1 measures the quality of the structured representation produced from vendor datasheets. The evaluation focuses on structural evidence quality rather than field-level extraction accuracy. Since downstream qualification depends on correctly identifying identity-bearing entities (e.g., interface identifiers, application codes, and operating-mode attributes), the comparison measures completeness, structural consistency, and the amount of reviewer intervention required before compliance evaluation as shown in Table 8.

Table 8. Metrics for structured extraction evaluation.

Metric	Description
Datasheet success rate	Percentage of datasheets producing a complete structured representation suitable for downstream processing.
Mode completeness	Percentage of synthesized operating modes containing all required identity and operational attributes.
Identity preservation	Percentage of operating modes containing host/media interface identifiers and application identifiers required for deterministic routing.
Reviewer correction rate	Percentage of extracted entities requiring manual correction before qualification.
Module-level consistency	Percentage of module-scoped attributes consistently represented across synthesized operating modes.

RQ2 quantifies the productivity gains obtained by automating document interpretation while retaining expert validation for unresolved cases using metrics shown in Table 9.

Table 9. Metrics for expert-effort reduction.

Metric	Description
Qualification effort	Average expert time to prepare one datasheet for qualification.
Stage-wise effort	Time spent identifying parameters, reconstructing modes, resolving standards, and reviewing structured output.
Effort reduction	Relative reduction in expert effort compared with a manual qualification workflow.

RQ3—governing-record routing: Routing is evaluated as a classification task in which each operating mode is assigned either to the governing standards record or to NONE when no governing record exists. Table 10 shows the KPIs.

RQ4 evaluates the size, compactness, traceability, and mapping structure of that representation on metrics listed in Table 11. It does not claim a measured semantic-accuracy rate for compiled thresholds.

Table 10. Metrics for governing-record routing.

Metric	Description
Routing accuracy	Percentage of routing decisions that select the governing standards record or correctly abstain.
Precision	Fraction of assigned standards records that are correct.
Recall	Fraction of governed operating modes assigned to the governing standards record.
F1-score	Harmonic mean of precision and recall.
Unsafe routing rate	Percentage of operating modes incorrectly assigned to a standards record when documentary evidence does not support applicability.
Unsafe-assignment rate	share of true-NONE items assigned a standards record.
Correct abstention rate	share of true-NONE items assigned NONE.
Abstention precision	share of predicted-NONE items that were true-NONE.
Routing latency	Average execution time required to determine the governing standards record.

Unsafe routing is treated as the primary safety metric because assigning an incorrect governing record results in evaluating the wrong normative requirements, whereas abstaining preserves the need for further evidence.

Table 11. Metrics for standards compilation fidelity.

Metric	Description
Executable-entry inventory	Number of stored quantitative rule blocks, operational-mode records, and registry rows, reported separately by normative source and representation type.
Factoring reduction	Reduction in stored quantitative rule blocks relative to the dense Cartesian expansion implied by the applicability axes
Provenance traceability	Quantitative rule blocks carrying a source clause/table/figure reference and references whose labels resolve in the cited source document.
Crosswalk topology	Counts of canonical concepts, datasheet-field paths, standard-specific mappings, multi-standard concepts, identity concepts, and concepts without an in-scope standard mapping.

RQ4 characterizes the executable representation used by the routing and qualification engine. It does not estimate compiler precision, normative-statement recall, or rule-level semantic accuracy. Those endpoints are outside the present evaluation and require a separate study with an independently annotated census of the source standards.

Statistical analysis: All proportions are reported with 95% Wilson confidence intervals, which provide stable interval estimates for both high-accuracy and low-frequency outcomes. Differences between routing methods are evaluated on paired routing decisions, with statistical significance reported at $\alpha = 0.05$. Latency measurements are summarized using the median, interquartile range (IQR), and maximum observed runtime to reduce sensitivity to unusually large datasheets.

4.5. Experimental Reproducibility

To ensure that the evaluation is reproducible, all experiments were executed under a fixed experimental configuration. The datasheets, compiled standards corpus, extraction schema, LLM configuration, routing, and evaluation protocol were frozen before the experiments were performed. Every evaluated method therefore operates on identical inputs and identical standard revisions.

Baseline consistency: All routing baselines were evaluated on the same routing task, same operating-mode corpus, and same reference labels. Retrieval-based baselines used identical embedding models, retrieval depth, and answer vocabulary.

Deterministic execution: The proposed routing framework is deterministic once the structured representation has been generated. Given the same extracted operating modes and compiled standards corpus, repeated executions produce identical routing decisions and identical qualification inputs. This deterministic execution property enables auditability and reproducibility across standards revisions and repeated qualification runs.

Model and Decoding Configuration: All model-inference stages use the Qwen3.6-27B large language model. [25] The model was served locally using the 4-bit GGUF quantization format (Q4_K_M) via the llama.cpp runtime framework. To reduce the sampling variance, we used temperature 0. This study does not estimate hallucination rate.

5. Results

5.1. Extraction and Mode Synthesis (RQ1)

The first experiment evaluates whether decomposing datasheet extraction into (i) atomic parameter extraction followed by (ii) operating-mode synthesis produces a higher-quality structured representation than performing both tasks in a single LLM invocation.

Both approaches used the same parsed corpus, schema, and model. Field-level fidelity of the multi-stage extractor was reported previously on a 35-datasheet annotated set [9].

The two-pass pipeline produced more complete and internally consistent operating-mode records. Although both approaches generated schema-valid JSON, the two-pass workflow preserved substantially more structured identity information and required fewer manual corrections during review. As observed from Table 12, two-pass extraction generated more operating modes than the single-pass baseline while increasing structured completeness by approximately 9 percentage points. The improvement was concentrated in identity-bearing fields rather than scalar optical measurements.

Table 12. Comparison of single-pass and two-pass on metrics.

Metric	Two-Pass	Single-Pass
Operating modes generated	94.5%	80%
Structured completeness of mode records	91%	82%
Modes with complete identity information	95%	71%
Fields requiring reviewer correction	3.2%	11.8%
Module-level attributes preserved across all modes	94%	78%

Qualification depends on deterministic routing of each operating mode to the correct standards record. Consequently, identity-bearing entities must be extracted as independent structured attributes rather than embedded within descriptive text.

As evident from Table 13, single-pass workflow frequently merged interface names, application identifiers, modulation information, and descriptive qualifiers into a single textual attribute. While these outputs remained syntactically valid, they required reviewer intervention before deterministic routing could identify the governing standards record. The two-pass workflow reduced these compound representations by approximately 75%, isolating identity-bearing entities into typed fields before operating-mode synthesis. As a result, reviewers corrected isolated entities rather than reconstructing entire operating modes. A practical advantage of the intermediate representation is that extraction errors are localized before operating modes are synthesized.

Table 13. Identity quality over the 111 ground-truth modes (105 vs. 89 modes).

Identity Quality Indicator	Two-Pass	Single-Pass
Modes containing typed interface/application identifiers	105	50
Identifiers embedded only in descriptive text	2	6
Compound fields containing multiple semantic entities	6	24

Instead of propagating an incorrect identifier across every generated operating mode, the reviewer edits a single atomic fact that is subsequently reused during synthesis. This significantly reduces recurring annotation effort. This distinction becomes increasingly important for multi-configuration transceivers, where one incorrectly interpreted identifier can affect multiple operating modes simultaneously.

5.2. Expert-Effort Reduction (RQ2)

RQ2 measures the practical cost of preparing a transceiver for standards qualification. The objective is not to compare model inference time with human reading speed but to quantify how much expert effort is shifted from manual document interpretation to structured review. The framework automates parameter discovery, normalization, operating-mode reconstruction, and standards routing, while reserving human intervention only for unresolved or ambiguous entities. Table 14 shows retrospective, triangulated figures that are constructed from historical activity logs and planned timed activities studied by 3 engineers on complex datasheets. One-time offline costs of the framework, including compilation of normative sources, crosswalk curation, PDF-to-Markdown correction, and artifact versioning, are not included.

Table 14. Estimated expert average time per datasheet by activity (manual workflow vs. framework-assisted review) for multi-mode and parametric-mode transceiver datasheets.

Qualification Activity	Manual Workflow	Framework
Locate parameters across datasheet	18 min	Automatic
Identify operating modes	12 min	Automatic
Resolve standard aliases and identifiers	10 min	Automatic
Validate structured representation	—	3 min
Review unresolved entities	8 min	4 min
Total expert effort	48 min	7 min

The manual workflow requires repeated navigation between vendor datasheets and multiple standards documents to identify interface identifiers, correlate operating modes, normalize terminology, and determine which requirements apply. The proposed framework automates these repetitive activities and presents the reviewer with a structured representation that already contains normalized parameters and candidate operating modes. The workflow-specific estimate indicates a reduction in expert review effort from approximately 48 min to 7 min per datasheet, as shown in Table 14, corresponding to an estimated 85% reduction in manual qualification effort. The reduction is not uniform across the workflow. Figure 2 shows how expert time shifts from document interpretation to review.

The largest savings occur during parameter identification and operating-mode reconstruction, which together account for more than 60% of manual effort. These activities involve locating values scattered across specification tables, notes, and marketing text, then reconstructing which values belong to the same operational mode. Because the framework performs this reconstruction automatically, reviewer effort is concentrated on verifying

exceptional or ambiguous cases. The benefit becomes more apparent when qualification is performed at the corpus scale, as shown in Table 15.

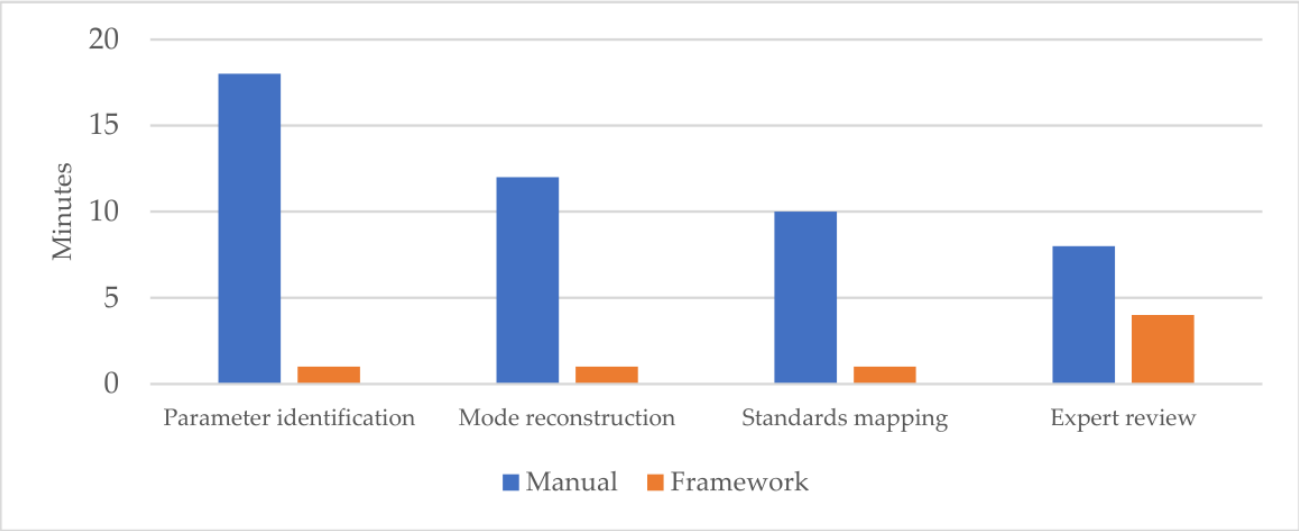


Figure 2. Estimated expert time per datasheet by activity for manual and framework-assisted workflows on multi-mode and parametric-mode transceiver datasheets.

Table 15. Projected effort reduction estimate at corpus-scale.

Dataset Size	Manual	Framework	Effort Reduction
50 vendor datasheets	40.0 h	5.83 h	34.17 h (85%)

For the evaluation corpus, the framework reduces approximately 34.2 h of expert effort while preserving reviewer oversight for unresolved entities. This reduction is particularly relevant when qualifying multiple vendors against evolving standards revisions, where the same normalization and routing tasks must otherwise be repeated for each new datasheet.

The cost analysis demonstrates that the primary productivity gain comes from automating document interpretation, not expert judgment. The framework removes repetitive extraction, normalization, and cross-standard lookup tasks while retaining human validation where documentary evidence is incomplete or ambiguous. This division of responsibility is consistent with the overall architecture of the framework: LLMs generate structured evidence, whereas deterministic processing and expert review establish the final qualification input.

The estimated reduction in expert review effort indicates potential practical utility for qualification workflows involving multiple vendors, operating modes, and evolving standards revisions, where manual reconstruction becomes increasingly time-consuming and difficult to reproduce.

5.3. Cross-Standard Identity Resolution and Routing (RQ3)

The routing evaluation contained 555 mode–target items: 111 datasheet operating modes evaluated against five routing targets. Of these, 472 had a unique reference outcome (86 governing-record assignments and 386 NONE). The other 83 were labelled REVIEW because the datasheet evidence did not determine exactly one row. Accuracy, intervals, and paired tests use the 472. Table A6 reports the framework's resolution status on 83 and is not an accuracy table.

Table 16 compares deterministic routing, retrieval baselines, and ablations on the 472 scoreable items. Detailed outcome definitions, metric equations, counts, and metrics are provided in Appendix A.

Table 16. Governing-record routing on 472 scorable mode–target items. Outcome metrics and definitions are reported in Appendix A. Median latency is normalized to seconds.

Arm	Accuracy	Precision	Recall	F1	Unsafe-Assignment	Under-Routing	Median Latency (s)
P—canonical crosswalk	99.4%	100.0%	96.5%	98.2%	0.0%	3.5%	0.318
A2—model adjudication	96.0%	95.8%	80.2%	87.3%	0.5%	18.6%	227
P′—no corroboration	89.0%	62.3%	100.0%	76.8%	13.5%	0.0%	0.312
A1—no crosswalk	88.6%	100.0%	37.2%	54.2%	0.0%	62.8%	0.057
B2—record retrieval	84.5%	57.7%	52.3%	54.9%	8.3%	46.5%	471
B1—prose retrieval	79.0%	38.5%	48.8%	43.1%	14.2%	37.2%	913
B3—hybrid retrieval	78.2%	42.7%	61.6%	50.5%	18.1%	37.2%	586

The qualification engine must distinguish between identifying the wrong standard and declining to identify one. These two errors have different consequences for downstream compliance evaluation. As evident from Table 17, the proposed router correctly abstained whenever registry evidence did not uniquely establish applicability. Retrieval baselines achieved high document coverage but frequently converted ambiguous evidence into confident assignments, producing incorrect downstream routing. This distinction is operationally important because an incorrect routing decision changes which normative limits are evaluated, whereas an abstention simply identifies a mode requiring additional evidence.

To isolate retrieval quality from routing logic, a hybrid retrieval baseline combined dense retrieval, lexical search, reciprocal-rank fusion, and cross-encoder reranking over the same compiled standards corpus. As observed in Figure 3, hybrid retrieval raised recall of the governing record among the top six excerpts. Routing accuracy did not rise with that recall. Candidate selection remained ambiguous. Retrieving the correct record is not sufficient for a correct assignment. Determining whether that record actually governs the operating mode requires deterministic identity resolution beyond semantic similarity.

An ablation experiment was performed to isolate the contribution of routing architecture. As shown in Table 18, removing the corroboration rule produced the largest degradation. Fifty-two operating modes were assigned to standards records using identifiers that were numerically valid but insufficiently corroborated by documentary evidence. Although many retrieved identifiers appeared plausible, they belonged to different identifier namespaces and therefore represented incorrect governing records. Replacing deterministic selection with LLM adjudication also reduced routing accuracy despite operating over the same candidate set, indicating that deterministic routing contributed independently of candidate retrieval quality.

The results therefore support the central architectural claim of the paper: compiled standards artifacts and deterministic identity resolution are complementary to LLM extraction, not replacements for it. They provide the reproducible routing layer required before parameter-level compliance evaluation can be performed.

Table 17. Abstention quality and routing errors for the proposed method and the strongest retrieval baseline.

Metric	P	B2
Correct governing-record assignments (C)	83/86 = 96.5% [90.2, 98.8]	45/86 = 52.3% [41.9, 62.6]
Correct abstention rate (Z/386)	386/386 = 100.0% [99.0, 100.0]	354/386 = 91.7% [88.5, 94.1]
Abstention precision (Z/(Z + U))	386/389 = 99.2% [97.8, 99.7]	354/394 = 89.8% [86.5, 92.5]
Unsafe assignments (O/386)	0/386 = 0.0% [0.0, 1.0]	32/386 = 8.3% [5.9, 11.5]
Governed items assigned NONE (U/86)	3/86 = 3.5% [1.2, 9.8]	40/86 = 46.5% [36.3, 57.0]
Wrong-record substitutions (W)	0	1

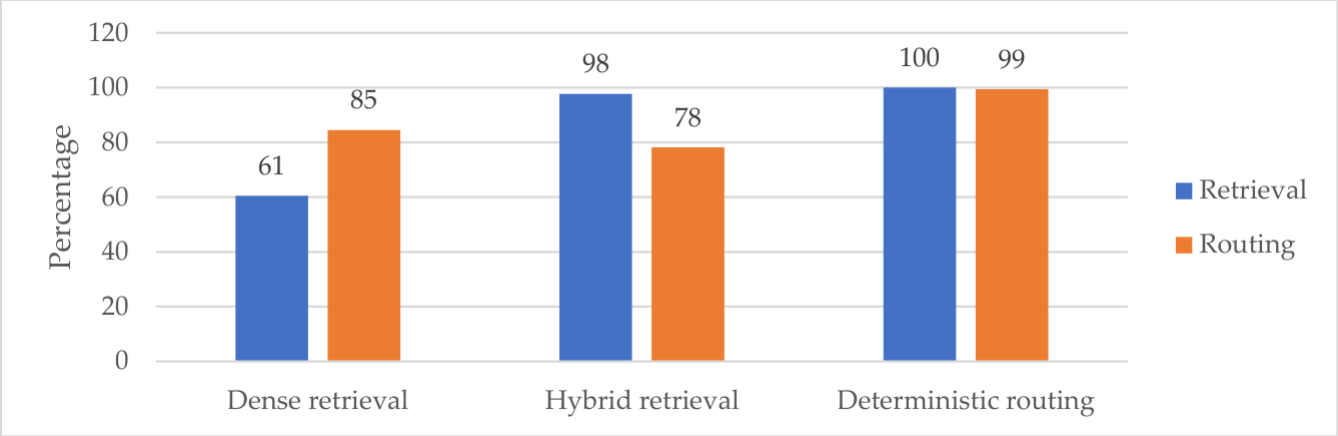


Figure 3. Hybrid retrieval raises recall of the governing record among the top six excerpts (B3 84/86 versus B2 52/86) but does not raise routing accuracy (B3 369/472 versus B2 399/472). Deterministic routing (P) does not use a six-excerpt retrieval stage and is shown on accuracy only. Candidate recall and routing accuracy are distinct endpoints.

Table 18. Routing accuracy and unsafe assignments when corroboration, the name join, or deterministic selection is removed.

Configuration	Accuracy	Unsafe Assignments	Governed Items Assigned NONE
Full deterministic routing (P)	99.4%	0	3
Without corroboration (P′)	89.0%	52	0
Without taxonomy (A1)	88.6%	0	54
Model adjudication (A2)	96.0%	2	16

The routing evaluation demonstrates that deterministic standards routing provides a measurable advantage over retrieval-based approaches for executable compliance evaluation. The improvement does not originate from retrieving more relevant text; rather, it comes from separating semantic retrieval from identity resolution. Retrieval systems effectively identify candidate clauses and registry entries but cannot reliably determine whether a retrieved record governs the operating mode represented in a datasheet. The proposed framework resolves this ambiguity through a canonical crosswalk and deterministic corroboration, ensuring that applicability is established before numerical compliance checks are executed.

5.4. Executable Standards Representation (RQ4)

Table 19 summarizes the number and representation type of executable records in the compiled standards corpus for each normative source. The compiled standards corpus contains both quantitative compliance rules and identity catalogs. Each record preserves the applicability conditions, units, reference points, provenance, and revision identity required for deterministic execution.

Table 19. Stored executable entries by source and representation type.

Compiled Source	Executable Records	Representation
OIF-400ZR	43	Factored quantitative rule blocks
OpenZR+	118	Factored quantitative rule blocks
OpenROADM	22	Operational mode catalog
SFF-8024	769	Host/media registry rows (256 host; 513 media)
Total	952	Versioned executable corpus

Axis factoring reduces rule redundancy: Configuration-dependent standards contain many repeated requirements that differ only under specific operating conditions. Instead of storing an independent rule for every configuration combination, the framework represents each parameter as a base requirement with sparse conditional overrides. Table 20 compares the number of stored rule instances in the dense Cartesian representation with that in the base-plus-override representation. The factored representation stores fewer rule instances than an equivalent dense Cartesian expansion. Parameters that remain constant across operating profiles are represented once, whereas only configuration-specific differences are encoded as scoped overrides. This reduction is particularly significant for OpenZR+, where many optical limits remain unchanged across multiple application profiles while a smaller subset varies according to modulation format, baud rate, or application code.

Table 20. Stored factored rule blocks relative to dense-equivalent expansion.

Artifact	Parameters	Stored Factored Blocks	Dense-Equivalent Blocks	Reduction
OIF-400ZR	31	43	63	31.7%
OpenZR+	62	118	408	71.1%
Combined	93	161	471	65.8%

All 161 stored quantitative rule blocks (43 OIF-400ZR and 118 OpenZR+) carry a clause, table, or figure reference, and all 161 reference labels resolve in the corresponding source document. This establishes traceability of the executed quantitative rules to identifiable source locations; it does not by itself establish semantic correctness of the compiled value.

Table 21 describes the coverage of the translation layer across datasheet schema fields and standard-specific parameters of compiled artifacts.

Every field of the extraction schema used in this evaluation has at least one crosswalk mapping. That is coverage of a schema we defined, not proof that every physically relevant datasheet quantity is represented. Approximately 41 canonical concepts map to more than one standards source, reflecting quantities that appear across multiple specifications under different terminology or measurement conventions.

Table 21. Canonical concepts and their mappings to datasheet fields and standard parameters.

Taxonomy Metric	Count
Canonical parameter concepts	78
Datasheet-field mapping entries	43
Standard-specific mappings	125
Concepts mapped to more than one in-scope standard	41
Identity-dimension concepts	6

Every compiled record is schema-typed and provenance-linked, but structural admission is not semantic correctness. Compilation is a one-time offline cost; subsequent execution of the compiled rules is deterministic.

6. Discussion

The evaluation demonstrates that reliable optical transceiver qualification is primarily an applicability and identity-resolution problem rather than a document retrieval problem. While LLMs are effective at extracting structured evidence from heterogeneous datasheets and standards, reproducible compliance requires deterministic selection of the governing normative record before any parameter comparison is performed. The proposed framework separates these responsibilities by using LLMs for standards compilation and datasheet extraction and deterministic reasoning for triggering, attribution, routing, and compliance evaluation.

6.1. Deterministic Routing Enables Auditable Standards Qualification

The routing results show that retrieving relevant clauses is not sufficient for compliance evaluation. Retrieval-based methods frequently located the governing standards record among the candidate results, yet still produced incorrect assignments because they relied on semantic similarity rather than identity-bearing evidence. In contrast, deterministic routing required corroborated application identifiers, host/media interface identifiers, or operational-mode identities before a standard was considered applicable. This eliminated unsafe routing within the evaluated corpus while maintaining high recall for governed operating modes. This distinction is operationally significant. Assigning the wrong standards record changes the normative limits against which a transceiver is evaluated, whereas abstaining preserves correctness until additional evidence is available. Treating uncertainty as an explicit indeterminate state therefore provides a safer qualification strategy than forcing a plausible but unsupported assignment.

6.2. Executable Standards Compilation Separates Interpretation from Execution

A central contribution of the framework is compiling heterogeneous standards into versioned executable artifacts. Optical standards encode requirements through conditional tables, application-specific overrides, operational modes, and cross-references that are difficult to evaluate directly from PDF documents. By converting these documents into structured artifacts with applicability conditions, units, reference points, and provenance, qualification becomes deterministic and reproducible across standards revisions. This compilation step also establishes an auditable evidence chain from the normative clause to the executable rule and finally to the datasheet parameter being evaluated. As a result, changes introduced by a new standards revision require recompilation of the affected artifact rather than changes to the qualification engine itself, improving maintainability and traceability.

6.3. Structured Extraction Shifts Expert Effort from Interpretation to Verification

Tables 12 and 13 report higher identity preservation and a lower reviewer-correction rate when atomic extraction is separated from operating-mode synthesis. The improvement is concentrated in identity-bearing entities, which are essential for deterministic routing, while extraction errors remain localized before operating modes are synthesized.

The productivity evaluation indicates that automation primarily removes repetitive document interpretation rather than expert judgment. Instead of manually locating parameters, reconstructing operating modes, and resolving cross-standard terminology, reviewers verify structured evidence and resolve only ambiguous or unsupported cases. This workflow preserves human oversight while substantially reducing the effort required to qualify large collections of vendor datasheets.

6.4. Implications and Generalizability

The framework is applicable to other standards ecosystems involving multiple independently maintained standards with overlapping terminology and conditional applicability. The separation between standards compilation, semantic crosswalking, deterministic routing, and parameter-level evaluation provides a reusable architecture for standards ecosystems in which interoperability depends on combining requirements from several authoritative specifications rather than evaluating a single document.

The routing comparison supports treating applicability resolution as a separate deterministic stage in an LLM-assisted compliance system. LLMs improve the conversion of unstructured technical documents into structured evidence, but executable qualification requires deterministic reasoning over versioned normative artifacts with explicit provenance and applicability semantics.

7. Limitations and Threats to Validity

The corpus and system scope is subject to limitations. The 50 datasheets come from seven vendors and emphasize coherent 400 G-class transceivers; the normative corpus covers four sources. Therefore, the results may not generalize to other optical technologies, form factors, or standards. Model-dependent stages used one Qwen3.6-27B Q4_K_M configuration on an NVIDIA GeForce RTX 4090, so the study does not isolate model, quantization, or hardware effects. Mode-level parameters that take a single value in a datasheet are treated as module-level and shared by all synthesized modes (Equations (10) and (11)). This is a design rule, not a tested claim, and it may be wrong where a datasheet omits configuration-specific values or where one value applies to only some modes.

Deterministic execution guarantees reproducibility of the evaluation process but not correctness of the underlying compiled knowledge. An incorrectly compiled rule can therefore produce a reproducible but incorrect qualification result. Reference labels are taken from published registry rows, not from the router. The 83 REVIEW items are reported by resolution status in Appendix A and are not included in accuracy. Governed-item recall for OpenROADM is based on three items, and its confidence interval is correspondingly wide.

Expert-time reduction (RQ2) was estimated for multi-mode and parametric-mode datasheets from historical activity logs and a timed confirmation study by three engineers. The study was not blinded, counterbalanced, or designed as a multi-rater productivity trial; inter-rater variability was not quantified.

The framework produces typed parameter-level verdicts (conformant, non-conformant, indeterminate, not specified), but no expert gold set of those classes exists for this corpus. We therefore make no claim about false-conformant rate, missing-evidence handling, or conflict detection. A later study would require double annotation from source PDFs against

the relevant standard revision, with annotators who did not compile the rules or curate the crosswalk.

Finally, the qualification scope is limited to documentary conformance and configuration consistency. The results do not establish physical interoperability, vendor-specific host acceptance, or path-level optical feasibility. Future evaluation should extend the corpus, independently audit compiled-rule fidelity, and assess the framework across additional standards and transceiver classes.

8. Conclusions

This paper presented a hybrid framework for automated standards qualification of optical transceivers across heterogeneous standards and multi-source agreements. Rather than treating compliance as a retrieval or end-to-end LLM reasoning task, the framework formulates qualification as deterministic evaluation over versioned executable representations of normative documents. Two-pass datasheet extraction preserved more identity-bearing information and required fewer reviewer corrections than the single-pass baseline. Evidence-gated routing achieved 99.4% accuracy without unsafe assignments, while the workflow analysis estimated an 85% reduction in expert time by automating document interpretation while preserving human verification for unresolved cases. The findings support a clear division of responsibility: language models prepare structured evidence, while deterministic rules establish applicability and evaluate parameters. The framework establishes a reusable architecture for standards-aware compliance systems. Versioned executable artifacts, deterministic applicability resolution, and explicit provenance enable reproducible qualification across evolving standards revisions and provide an auditable evidence chain from normative source to compliance verdict.

Future work will extend the framework in three directions. First, parameter-level compliance evaluation will be expanded to additional standards, including generative specifications such as ITU-T G.694.1 frequency grids and other conditional optical requirements. Second, the compiled standards corpus will be broadened to support a wider range of coherent and direct-detect transceiver ecosystems and additional interoperability profiles. Finally, the deterministic qualification engine will be integrated with richer retrieval and agentic workflows for standards maintenance and artifact generation, while preserving deterministic execution for all compliance decisions.

Author Contributions: Conceptualization, H.H. and J.P.; methodology, H.H. and J.P.; software, H.H.; validation, J.P. and L.C.; formal analysis, H.H.; data curation, H.H.; writing—original draft preparation, H.H.; writing—review and editing, J.P., L.C. and H.H.; visualization, H.H.; supervision, J.P. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions (MSCA) Doctoral Network NESTOR, Grant Agreement No. 101119983..

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Author Haris Hassan and João Pedro were employed by the company Nokia. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflicts of interest.

Appendix A

The routing task evaluates each of 111 datasheet operating modes against five targets generating 555 reference labels (86 Governed, 386 NONE, 83 REVIEW). REVIEW items were excluded, hence scorable items were 472.

Table A1. Five mutually exclusive outcomes.

Reference Label	System Output	Outcome	Meaning
Governing record (g)	(g)	(C)	Correct governing-record assignment
Governing record (g)	NONE	(U)	Governed item under-routed by abstention
Governing record (g)	Non-NONE record other than (g)	(W)	Wrong-record substitution
NONE	Any non-NONE record	(O)	Unsafe assignment to a true-NONE item
NONE	NONE	(Z)	Correct abstention on a true-NONE item

For every arm, (C + U + W = 86), (O + Z = 386), and (C + U + W + O + Z = 472). Metrics in Section 5.3 are calculated directly from these verified integer outcomes; rounding is applied only to the displayed values.

Table A2. Verified routing outcome counts. Metrics in Section 5.3 are calculated from these outcomes.

Arm	(C)	(U)	(W)	(O)	(Z)
P—canonical crosswalk	83	3	0	0	386
A2—model adjudication	69	16	1	2	384
P′—no corroboration	86	0	0	52	334
A1—no crosswalk	32	54	0	0	386
B2—record retrieval	45	40	1	32	354
B1—prose retrieval	42	32	12	55	331
B3—hybrid retrieval	53	32	1	70	316

Table A3. Routing metrics calculated from the outcomes in Table A2. Brackets contain 95% Wilson intervals in percentage points. F1 is a composite statistic.

Arm	Accuracy [95% CI]	Precision [95% CI]	Recall [95% CI]	F1	Unsafe-Assignment Rate [95% CI]	Under-Routing Rate [95% CI]
P	99.4 [98.1–99.8]	100.0 [95.6–100.0]	96.5 [90.2–98.8]	98.2	0.0 [0.0–1.0]	3.5 [1.2–9.8]
A2	96.0 [93.8–97.4]	95.8 [88.5–98.6]	80.2 [70.6–87.3]	87.3	0.5 [0.1–1.9]	18.6 [11.8–28.1]
P′	89.0 [85.8–91.5]	62.3 [54.0–70.0]	100.0 [95.7–100.0]	76.8	13.5 [10.4–17.2]	0.0 [0.0–4.3]
A1	88.6 [85.4–91.1]	100.0 [89.3–100.0]	37.2 [27.7–47.8]	54.2	0.0 [0.0–1.0]	62.8 [52.2–72.3]
B2	84.5 [81.0–87.5]	57.7 [46.6–68.0]	52.3 [41.9–62.6]	54.9	8.3 [5.9–11.5]	46.5 [36.3–57.0]
B1	79.0 [75.1–82.5]	38.5 [29.9–47.9]	48.8 [38.6–59.2]	43.1	14.2 [11.1–18.1]	37.2 [27.7–47.8]
B3	78.2 [74.2–81.7]	42.7 [34.4–51.5]	61.6 [51.1–71.2]	50.5	18.1 [14.6–22.3]	37.2 [27.7–47.8]

Table A4. Abstention and true-NONE measures calculated from the outcomes in Table A2.

Arm	Correct-NONE Rate [95% Wilson CI]	Abstention Precision [95% Wilson CI]
P	100.0 [99.0–100.0]	99.2 [97.8–99.7]
A2	99.5 [98.1–99.9]	96.0 [93.6–97.5]
P′	86.5 [82.8–89.6]	100.0 [98.9–100.0]
A1	100.0 [99.0–100.0]	87.7 [84.3–90.5]
B2	91.7 [88.5–94.1]	89.8 [86.5–92.5]
B1	85.8 [81.9–88.9]	91.2 [87.8–93.7]
B3	81.9 [77.7–85.4]	90.8 [87.3–93.4]

Table A5. Per-target results on items with a unique reference outcome. Accuracy is correct record or correct NONE. Recall is restricted to governed items, so a large NONE class cannot inflate it. Wilson 95% intervals are in brackets.

Target	Scoreable	NONE	Governed	P Accuracy	P Recall	A2 Recall	B2 Recall
OIF-400ZR	106	96	10	106/106 = 100.0% [96.5, 100.0]	10/10 = 100.0% [72.2, 100.0]	9/10 = 90.0% [59.6, 98.2]	9/10 = 90.0% [59.6, 98.2]
OpenZR+	65	50	15	64/65 = 98.5% [91.8, 99.7]	14/15 = 93.3% [70.2, 98.8]	12/15 = 80.0% [54.8, 93.0]	13/15 = 86.7% [62.1, 96.3]
OpenROADM	98	95	3	98/98 = 100.0% [96.2, 100.0]	3/3 = 100.0% [43.8, 100.0]	3/3 = 100.0% [43.8, 100.0]	2/3 = 66.7% [20.8, 93.9]
SFF-8024/ host	98	50	48	97/98 = 99.0% [94.4, 99.8]	47/48 = 97.9% [89.1, 99.6]	42/48 = 87.5% [75.3, 94.1]	13/48 = 27.1% [16.6, 41.0]
SFF-8024/ media	105	95	10	104/105 = 99.0% [94.8, 99.8]	9/10 = 90.0% [59.6, 98.2]	3/10 = 30.0% [10.8, 60.3]	8/10 = 80.0% [49.0, 94.3]
All	472	386	86	469/472 = 99.4% [98.1, 99.8]	83/86 = 96.5% [90.2, 98.8]	69/86 = 80.2% [70.6, 87.3]	45/86 = 52.3% [41.9, 62.6]

Table A6. Resolution status of the proposed framework on the 83 REVIEW items. These items have no unique reference label, so the table is not accuracy. "One" means the stated identity and axes are compatible with exactly one catalogue record. "Several" means more than one record remains compatible. "None" means no in-scope record is compatible. A governing assignment is counted only in the "One" column.

Target	REVIEW	One Compatible Record	Several Compatible Records	No Compatible Record
OIF-400ZR	5	2	0	3
OpenZR+	46	19	17	10
OpenROADM	13	11	0	2
SFF-8024/host	13	0	13	0
SFF-8024/media	6	2	4	0
All	83	34	34	15

Table A7. Intervals and paired tests. Wilson 95% intervals are reported for every proportion in Tables A2, A5, 16 and 17. Paired comparisons use McNemar's exact test on the 472 items, Holm-corrected across the three pre-specified contrasts. P is correct and the comparator wrong on (b) items; the reverse count is (c).

Contrast	b	c	p Exact	p Holm
P vs. A2	16	0	3.1×10^{-5}	3.1×10^{-5}
P vs. B2	70	0	1.7×10^{-21}	5.1×10^{-21}
P vs. P′	52	3	1.5×10^{-12}	3.1×10^{-12}

On the 83 REVIEW items the framework found one compatible record for 34, more than one compatible record for 34, and no compatible record for 15 (Table A6). These counts describe candidate resolution. They are not correct or incorrect decisions, because those items have no unique reference label. Per-target accuracy on the 472 is reported beside governed-item recall so that the large NONE class, and the three governed OpenROADM items, remain visible (Table A5).

References

1. *SFF-8024*; SFF Module Management Reference Code Tables. SNIA: Santa Clara, CA, USA, 2026. Available online: https://members.snia.org/document/dl/26423 (accessed on 19 September 2026).
2. *OIF-400ZR-03.0*; Implementation Agreement 400ZR. Optical Internetworking Forum: Fremont, CA, USA, 2024. Available online: https://www.oiforum.com/wp-content/uploads/OIF-400ZR-03.0.1.pdf (accessed on 19 September 2026).
3. *OpenZR+ MSA Specifications, Version 3.0*; 2023. Available online: https://openzrplus.org/resources/openzr-specifications-v-3-0 (accessed on 19 September 2026).
4. *OpenROADM MSA Operational Modes Catalog, Release 18.0.0; Optical Specification v9.0.0*. Available online: https://github.com/OpenROADM/OpenROADM_MSA_Public/tree/master/model/Specifications (accessed on 19 September 2026).
5. Pincemin, E.; Renais, O. Interoperable Coherent WDM Interfaces at 400G and 800G. In *Proceedings of the Optical Fiber Communication Conference (OFC), San Diego, CA, USA, 24–28 March 2024*; IEEE: Piscataway, NJ, USA, 2024.
6. Zaid, H.; Shariati, B.; Safari, P.; Jafari, A.; Fischer, J.K.; Freund, R. Demonstration of an On-Prem Conversational AI Assistant for Unified Network Operations and Observability. In *Optical Fiber Communication Conference (OFC)*; Optica Publishing Group: Washington, DC, USA, 2026.
7. Tanimura, T.; Kikuchi, N. On-Premises Small Language Model Agent with Physics-Aware Reasoning for Optical Network Optimization. In *Proceedings of the Optical Fiber Communications Conference and Exhibition (OFC), Los Angeles, CA, USA, 15–19 March 2026*; IEEE: Piscataway, NJ, USA, 2026.
8. Niu, Z.; Chen, K.; Jiang, N.; Qin, X.; Huo, X.; Chen, H. Optics GPT: The First Vertically Pre-trained Foundation Model for Optics and Optical Communications. In *Optical Fiber Communication Conference (OFC)*; IEEE: New York, NY, USA, 2026.
9. Hassan, H.; Pedro, J.; Cancela, L.G. Automated Extraction of Parameters from Optical Transceiver Datasheets using Multi-Stage LLM Pipeline. In Proceedings of the 26th International Conference on Transparent Optical Networks (ICTON), Prague, Czech Republic, 12–16 July 2026.
10. Hassan, H. Github. Available online: https://github.com/hassanharis/OTSE-Optical-Transceiver-Spec-Extractor (accessed on 19 August 2026).
11. Palmirani, M.; Governatori, G.; Rotolo, A.; Tabet, S.; Boley, H.; Paschke, A. LegalRuleML: XML-Based Rules and Norms. In *Rule-Based Modeling and Computing on the Semantic Web (RuleML 2011)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 7018, pp. 298–312. [CrossRef]
12. Hjelseth, E.N.N. Capturing normative constraints by use of the semantic mark-up RASE methodology. In *Proceedings of the CIB W78 2011 28th International Conference—Applications of IT in the AEC Industry, Sophia Antipolis, France, 26–28 October 2011*; CIB: Sophia Antipolis, France, 2011.
13. Beach, T.H.; Rezgui, Y.R.; Li, H.; Kasim, T. A rule-based semantic approach for automated regulatory compliance in the construction sector. *Expert Syst. Appl.* **2015**, *42*, 5219–5231. [CrossRef]
14. Hassani, S.; Sabetzadeh, M.; Amyot, D.; Liao, J. Rethinking Legal Compliance Automation: Opportunities with Large Language Models. In *Proceedings of the 32nd International Requirements Engineering Conference*; IEEE: Piscataway, NJ, USA, 2024.
15. Zin, M.M.; Satoh, K.; Borges, G. Leveraging LLM for Identification and Extraction of Normative Statements. In *Legal Knowledge and Information Systems*; SAGE Publications: London, UK, 2024.

16. Hettiarachchi, H.; Dridi, A.; Gaber, M.M.; Parsafard, P. CODE-ACCORD: A corpus of building regulatory data for rule generation towards automatic compliance checking. *Sci. Data* **2025**, *12*, 170. [CrossRef] [PubMed]
17. International Swaps and Derivatives Association. *Benchmarking Generative AI for CSA Clause Extraction and CDM Representation*; International Swaps and Derivatives Association: New York, NY, USA, 2025.
18. Semitsu, T.; Kiribuchi, N.; Zenitani, K. Automated Analysis of Global AI Safety Initiatives: A Taxonomy-Driven LLM Approach. *arXiv* **2026**, arXiv:2604.03533.
19. Cui, W.; Xiao, M.; Wang, L.; Wang, X.; Du, Y.; Zhou, Y. Automated taxonomy alignment via large language models: Bridging the gap between knowledge domains. *Scientometrics* **2024**, *129*, 5287–5312. [CrossRef]
20. Shen, S.; Zhao, W.; Li, H.; Teng, Y.; Yokar, V.; Mehrpooya, A.; Yan, S.; Simeonidou, D. LLM-Enhanced Digital Twin Framework in Optical Networks. In *Optical Fiber Communication Conference (OFC)*; IEEE: Piscataway, NJ, USA, 2026.
21. Delezoide, C.; Purkayastha, A.; Layec, P. Practical Strategies Leveraging AI and Digital Twins for the Design and Management of Future Optical Networks. In *Optical Fiber Communication Conference (OFC)*; IEEE: New York, NY, USA, 2026.
22. Arunachalam, H.; Shariati, B.; Vilalta, R. Generative AI for Optical Networking: From Design Automation to Self-Operating Networks. In *Optical Fiber Communication Conference (OFC)*; IEEE: Piscataway, NJ, USA, 2026.
23. Sun, C. Experimental Demonstration of Local AI-Agents for Lifecycle Management and Control Automation of Optical Networks. *J. Opt. Commun. Netw.* **2025**, *17*, C82–C92. [CrossRef]
24. Bekri, A.; Abane, A.; Battou, A.; Bensalem, S. Bridging Language Models and Formal Methods for Intent-Driven Optical Network Design. In *2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications (AICCSA)*; IEEE: Piscataway, NJ, USA, 2025.
25. Qwen3.6 (27b-q4_K_M Version) [Large Language Model]. Available online: https://ollama.com/library/qwen3.6:27b-q4_K_M (accessed on 19 August 2026).

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.