



DRAFT FOR COMMENTING

OIML SMART PROGRAM · SYSTEM DESCRIPTION AND MEASURED EVALUATION

OIML SMART AI: a grounded question-answering service over legal-metrology publications

The knowledge staircase, the answer contract, execution semantics, the input data kinds and their enrichment, and platform-portable deployment — measured over ~900 OIML publications.



2026-09-10

ABSTRACT

Legal-metrology knowledge is locked in documents: a practitioner's question ("what is the minimum number of verification intervals for a class B load cell?") is answered by a cell of a table in a specific edition of a specific publication, and neither keyword search nor a language model's memory can be trusted to produce it. We describe OIML SMART AI, a production question-answering service over the OIML publication corpus that treats corpus structure as the decisive variable: a measured comparison across six representations of identical content shows answer capability rising from 2/18 probes (plain text) to 13/18 (executable machine models), and the full serving stack adds a further three rungs to 17–18/18. The system enforces an answer contract — clause-exact citations, verbatim normative values, tables and figures rendered as producer-owned typed objects, conformance answered by executing machine-checkable constraints — with deterministic post-generation checks. We report the layered architecture and its platform-portable techniques, the input data kinds and enrichment pipeline that feed it, the answer contract, corpus-integrity operations learned from production incidents, and the economics of a reference deployment on open-weight models.

§ 1 • INTRODUCTION

Documents do not answer questions

The OIML publishes the Recommendations, Documents and Basic publications that define legal metrology worldwide. Finding the clause that governs a question has historically required already knowing the document. A conventional keyword search fails unless the question echoes the document's wording; a conventional large language model answers fluently from parametric memory, which for normative content is worse than failing — **an ungrounded answer about a maximum permissible error is indistinguishable from a grounded one until an instrument fails verification.**

Legal metrology adds constraints that generic retrieval systems do not carry:

- **Editions supersede.** OIML R 60:2021 is three parts; the 2017 consolidated edition is superseded but topically near-identical. A system that mixes editions silently answers from law that no longer applies.
- **Values are normative and tabular.** Class limits, MPE tiers and verification bounds live in tables; flattening them into prose destroys the row×column geometry that makes a cell answerable, and paraphrasing "5 000" as "about five thousand" is a defect, not a style choice.
- **Some answers are computations.** "Is a proposed $D_{\max}$ of 26 000 v acceptable for $E_{\max}$ 30 000 v?" is decided by a constraint the document defines and never prints as an answer.
- **Provenance is the product.** An answer without the exact publication, edition and clause is unusable in a regulatory context.

This paper describes the system built against those constraints: its knowledge representation thesis (§2), layered architecture (§3), the input data kinds and enrichment pipeline that feed it (§4), the answer contract (§5) and execution semantics (§6), measurement methodology and results (§7), corpus-integrity operations (§8), and deployment modes with economics (§9).

The annealment staircase

The system's central, measured finding: **the representation of a corpus determines the ceiling of what questions it can answer** — not incrementally but structurally. A representation either carries a primitive (typed tables, vocabulary registers, machine-checkable constraints) or it does not; questions requiring that primitive are answerable or not. We index the same corpus — the OIML R 60 family — under six representations of increasing structural *annealment* and measure each with 18 witness-graded capability probes spanning a ten-rung ladder from LOCATE (find the clause) to INSTANCE & PROCESS (ordered test sequences with contamination semantics):

Representation	Carries	Probes answered
A — plain text	words	2/18
B — AsciiDoc markup	words + markup	3/18
C — typed units (MKO)	tables/formulas/figures as objects; cross-references as links	6/18
D — executable model (Primmel)	constraints, calculations, sequences, terms (P1–P10)	13/18
E — ablation (D without typing)	the causal control	12/18
F — composed (C + D)	document + model	14/18
Full serving stack	any representation + the mechanisms of §3	17–18/18

Three results deserve emphasis. First, the machine-model lane is **alone able to answer rungs L8–L9** (computation and process semantics) — no document representation contains constraint or sequence objects at all. Second, the serving stack is worth **+3 rungs over the best single representation with identical content**: representation determines what is retrievable; the pipeline determines whether it is used. Third, the universal failure mode across all lanes — the colloquial-to-defined-term bridge ("my output keeps drifting" → *durability*, VIML 5.15) — is closed not by representation but by a serving-side vocabulary binding, a finding with direct consequences for standards producers: unit-level vocabulary registers would move that capability into the corpus itself.

The thesis in one line. Anneal knowledge into structure a machine can traverse, and the questions a corpus can answer expand structurally — from lookup, through geometry, to computation.

A stage-registry pipeline on edge infrastructure

The architecture is defined by its layers, not by any platform. Retrieval is composed as an ordered registry of stages; adding a mechanism is a new stage module, never a modification of existing ones (open/closed as an engineering constraint, versioned in the pipeline specification). The layers depend on four managed capabilities — a vector index, a key-value cache, a relational store, an object store — plus any runtime that can execute the registry; every major cloud offers them, and a self-hosted Kubernetes deployment is equally valid. The reference deployment described throughout is one instantiation (edge functions + managed vector/KV/SQL/object services, open-weight models served on-platform); §9 discusses deployment modes.

Two indexes are maintained, structurally isolated: the public OIML corpus and a members' ISO/IEC corpus. Isolation is architectural — no binding, credential or route from the public tier reaches the internal index — so an audience permission is enforced by construction, not by filtering (§6's leakage probes therefore test implementation bugs, not policy).

3.1 Query understanding and deterministic scope

One small model reads every question for language, named publication, edition, process intent, decomposition and hypothetical-answer variants — meaning, never string matching. Where the question *names a publication in its own text*, scope is applied deterministically by parsing the question, not by trusting the model's per-call extraction (measured: leaving it to the model made "What is OIML D 29?" a coin flip between a correct scoped retrieval and a refusal citing the wrong family).

3.2 Hybrid retrieval with structural ranking

Dense vectors search alongside a full-corpus BM25 scan, RRF-fused, with a terminology graph lane, hypothetical-answer keys, and multi-phrasing lanes running concurrently. Citation-shaped questions take their own graph lane: what each edition's bibliography cites is extracted as graph edges, and the answer receives the cited-standards list as authoritative structured data alongside the retrieved bibliography passage that grounds it verbatim. Every chunk carries a model-written contextual preamble stating where it sits in its document (98.7% coverage, written once at index time — its quality persists into every future retrieval). The corpus *is* a tree: a hit's score blends its ancestors' and descendants'; evidence is presented in document reading order; same-chain near-duplicates collapse. Superseded editions demote family-relative — full-strength under either signal: the successor edition is present in the pool, or the chunk's own status says superseded — while remaining citable when only they carry content. A registry-driven cover stage guarantees the *active* edition can enter a pool that drew only stale editions (the measured failure: "which R 60 edition applies this year?" answered 2017 under unlucky rerank draws).

3.3 Vocabulary binding

Everyday words rarely match defined terms. A glossary of 8.8k concepts links the question to candidate terms (dense candidates + cross-encoder rerank); the answer model adjudicates and leads with the corpus's own term, quoting its defining publication — with the time-scale discriminator that separates creep (minutes under load) from durability (months of service).

3.4 Dataset scope and estate permissions

Three databases — OIML Publications, OIML SMART Models, ISO/IEC Conformity Assessment — are selectable per question; the server intersects the requested scope with session permissions (the ISO/IEC corpus requires the `ai-preview` estate permission; membership alone is not the bar). A corpus-scope stage filters the assembled candidate pool after every refill path, so disabling a database disables it at the hit level.

What enters the system, and what it becomes

The pipeline is fed by producers, not files: each input kind has a owner and a derivation, and every stage of enrichment is recorded and replayable.

Input kind	Origin	What the system derives
Authoried Metanorma sources	hand-curated publications (XML + sections + clause tree)	clause-anchored chunks along authored boundaries; typed units (tables, formulas, figures with assets); cross-reference edges
OCR-derived publications	~880 scanned documents re-authored to Metanorma	same chunk + anchor discipline, with a quality tier marking provenance; shells (failed OCR) flagged, never served as content
Bibliography records	~5 700 structured citations with supersession edges	the edition registry: current/superseded derived from successor links (never trusting a disagreeing status field); the citation graph
Terminology datasets	glossaries of ~8 800 concepts with defining publications	the vocabulary lane: everyday-word to defined-term bridging with model adjudication
Recommendation models	machine-checkable packages (requirements, constraints, test sequences)	the model plane: typed model chunks for retrieval AND the node store the execution engine of §6 evaluates

Enrichment is an explicit, replayable stage: every chunk gains a model-written **contextual preamble** situating it in its document (where it sits, what it governs) — written once at index time, embedded with the text, so its quality persists into every future retrieval; the preambles live in the index with a durable record beside it, because full restores overwrite them and must replay (§8). Chunk identity is content-addressed and every derivation writes its own artifact — three derivations sharing one file path once silently overwrote each other, a class of bug the artifact discipline now prevents. All chunk writes cross a single typed gateway (a schema-checked adapter with target gating): lane corpora structurally cannot enter a production index, and a hand-rolled metadata dictionary cannot bypass validation.

Claims are checkable

Every answer passes a contract that generation cannot waive:

- **Citations** name the publication, edition and clause on every claim.
- **Normative values are verbatim** — mechanically verified as quote-containment in the served passages, with one corrected regeneration when a check fails.
- **Typed objects never pass through the model.** Tables, formulas and figures are referenced symbolically (`[[u:...]]`) and resolved server-side to the producer's payload; the model emits a reference, the interface renders the exact object. A model cannot corrupt a table it never transcribes.
- **Figures are read, not paraphrased.** A pinned figure's pixels ride the generation call, so labels that exist only in the drawing (L7's pixel-only probes) are read from the drawing; assets are verified readable by vision pipelines (a vector-sourced raster drawing black on transparent alpha flattens to a solid black rectangle in a vision model — a defect class browsers hide by compositing on white).

Citations are doors, not labels. Every clean-corpus publication is rendered to a self-contained HTML carrying the corpus's own clause anchors; a citation opens the original document at the cited clause — highlighted, neighbors visible — inside the answer. The reader verifies the quote against the document itself, not against our snippet. Provable absence completes the contract: "does R 60 constrain packaging?" enumerates the standard's entire model plane and returns a certificate — *absent, N nodes enumerated, 0 matches, scope named* — never a bare refusal. Any answer is independently checkable through a verification endpoint (quote containment, object-reference resolution, judged faithfulness labeled as judged).

Conformance by computation

When a question names a machine-checkable object, the service executes it. The question's values bind to the rule's own parameters; a deterministic evaluator runs the check (OCL boolean expressions over the model plane, threshold limits); the verdict — **pass**, or **the standard's own violation wording with the arithmetic shown**, or **void naming the missing parameters** — is attached as data the answer model must present faithfully. Counterfactuals are free: hypothetical values are just values, and table lookups run inverted ("which accuracy class for $n_{LC} = 3\,000$?"). Ground truth for these rungs is computed offline from the same model — no judge in the measurement loop. This is the capability no text representation can follow at any scale, and the direction the ladder points: the corpus as a model, queried by execution.

One command, honest exit codes

Every change ships through the same gate: a 38-case golden suite ×3 (witness-span containment grading: doc-level, definitions, table values, refusals, filters, French and German, leakage probes, verdicts) and the 18-probe capability battery ×6, any failed run failing the command. Run logs are evidence — stored durably, never in ephemeral directories. Current standing from the definitive post-restore run: **golden 37–38/38 per run (97–100%)**; capability battery **mode 18/18** with two known single-probe variances (a cross-standard retrieval probe; a normative-date phrasing wobble) — reported as ranges, never peaks. The same gate settled the effort question. Sweeping the answer model's reasoning effort on our own corpus found a coupling the model reports omit: elevated effort with a flat output budget is *worse* than no effort at all — 5/38 golden, truncating mid-answer — while the same effort with the budget coupled to it passes 36–38/38 across three consecutive runs. Reasoning that cannot finish its answer spends tokens to lose accuracy; the deployed Thorough lane carries both, by measurement. Leakage probes gate every promotion: questions answerable only from the internal ISO/IEC corpus must return "no information" from the public tier, and do — the architecture cannot leak (separate indexes, no binding from the public worker), so the probes test implementation bugs.

A retrieval service that never audits its own index serves ghosts. Production incidents hardened the system into an explicit operations loop, each step a durable script:

- **Reconciliation.** Vector upserts never delete; the first census found 49 553 live vectors against 31 512 canonical — 18 041 strays from every prior re-chunking — deleted in place, with the canonical set defined as three derivations (parse, retrieval-plane export, typed-unit projection) that must never overwrite one another's artifacts.
- **Enrichment replay.** Contextual preambles live only in the index; a full restore overwrites them with raw text (measured: 0% enriched after one restore, identifier queries failing first). The durable record the enrichment driver writes enables a zero-generation replay of 29 646 chunks.
- **Asset readability sweeps.** The black-alpha defect class of §4, detected and repaired in place with original backups.
- **Cache generation.** Answer caches are namespaced by deploy version and a corpus-generation stamp; corpus surgery bumps the stamp and stale entries miss. Request-scoped context (database selection, personalized memory files) salts the cache keys — a memory-flavored answer can never serve a memory-less ask.

Open weights at the edge

9.1 Portable by construction

Deployment is an implementation detail of the reference service, not a property of the techniques. The layers require only the four managed capabilities of §3 and a model endpoint; adopters have run the same registry over self-hosted vector databases and containerized inference. What transfers intact: the stage registry and its thresholds, the chunk wire contract, the adapter gateway, the enrichment/replay loop, the measurement gate, and the answer contract — the entire reasoning of the system is platform-neutral code and data.

9.2 The reference instantiation

One deployment — edge functions, managed vector/KV/SQL/object stores, open-weight models served on-platform — demonstrates the economics: a ~320B-parameter natively multimodal answering model at **≈\$0.0012 per answer**; an anonymous tier on a 3B-active MoE at ≈\$0.0003; a ~\$0.15 embedding pass for the whole corpus; no proprietary model APIs and no data leaving the estate. One-time quality work (enrichment, graph build, golden-set drafting) runs on the best available model regardless of price — bounded by corpus size, its quality persists into every future answer.

Members additionally hold personalized memory files (per-account context documents, selected per question like databases), projects that let a group of chats share one evolving memory, and per-conversation database scoping — every selected context salts the answer cache, so personalization can never serve one member another's cached answer.

What the staircase enables

For the metrology community: the corpus becomes queryable in the ways practitioners actually ask — colloquially, across editions, into tables, and against configurations the documents define but never print. A regulator drafts against the current edition with the delta to its predecessor a clause-lookup away; a laboratory checks a configuration by execution instead of reading; a committee asks impact questions over the requirement–test–formula graph. For standards publishers, the staircase is the lesson this system was built to measure: structure added to the corpus is capability added to every consumer of it — and the highest rungs (verdicts, proven absence, self-verification) exist only when the publication itself is a model.

Artifacts. Live service: ai.oimlsmart.org · mechanism reference and ladder elaboration: the project documentation set ([docs/sota-mechanisms.md](#), [docs/knowledge-annealment.md](#)) · measured comparison study: [docs/paper-annealment-comparison.md](#) · capability panel: ai.oimlsmart.org/annealment.

OIML SMART AI · system description and measured evaluation · September 2026 · drafted for the OIML SMART programme