# ELOG-RAG

*Deep answers from measurement-heavy logbooks.*

Finding a needle in the needle stack.

Alexander Warth · SC Magnets & Testing · GSI/FAIR AI Workshop, 31 August 2026

Supported by the GSI Innovation Fund · External industry pilot partner · Runs fully on-premise

# Can a system understand this?

**#8796 · 17.08.2026 · QDM-040 · cooldown run CD02 · "Bipolar Cold Run 2"**

```
Performed: %SH_Bipolar%, %SV_Bipolar%, %CS_Bipolar%
** CS ** Bipolar Stability (10 minutes)
1. [16:01:37], -250 A -> 0.000 s -> 250 A -> 0 A -> -250 A,
   +1225.0 A/s, -1225.0 A/s --> CS Quench: [No];
** SH + SV ** Bipolar Stability Validation (10 minutes)
Green. [16:14:28], -100 A -> 250 A -> 0 A -> -100 A, +/- 857.5 A/s
   --> SH Quench: [Yes @ 16:14:34], SV Quench: [No]; #Mistake
Green. [16:17:07], -100 A -> 250 A -> 0 A -> -100 A, +/- 857.5 A/s
   --> SH Quench: [No], SV Quench: [No];
...
```

**ASK-AI** *"Did SH quench above 800 A/s in cooldown 02 of QDM-040?"*

***Answer:***

….Entry 8796 (Bipolar Stability, 17.08.26): SH + SV Bipolar Stability Validation at 857.5 A/s – one instance showed `SH Quench: [Yes @ 16:14:34]`, ***but this was flagged as a "#Mistake" and subsequent runs at the same rate were stable***….

### Dense by construction

Timestamped ramp sequences, currents, ramp rates, quench yes/no with time, operator remarks - several facts per line, in a house notation that never appears in any training corpus.

**2,702** entries carry explicit measurements with units

**79** authors over twelve years – notation drifts

**97 % EN · 3 % DE** bilingual, sometimes within one entry

*The answer below is the system's, verbatim: it finds the entry, reads the quench, and reads the operator's #Mistake flag right after it.*

# 8,712 of those. Twelve years of them.

**8,712**

**logbook entries**

Series Test Facility ELOG, July 2014 to today: 157 modules, every ramp and quench.

**12**

**years**

Reaches back well before the current campaigns. Vocabulary and authors change.

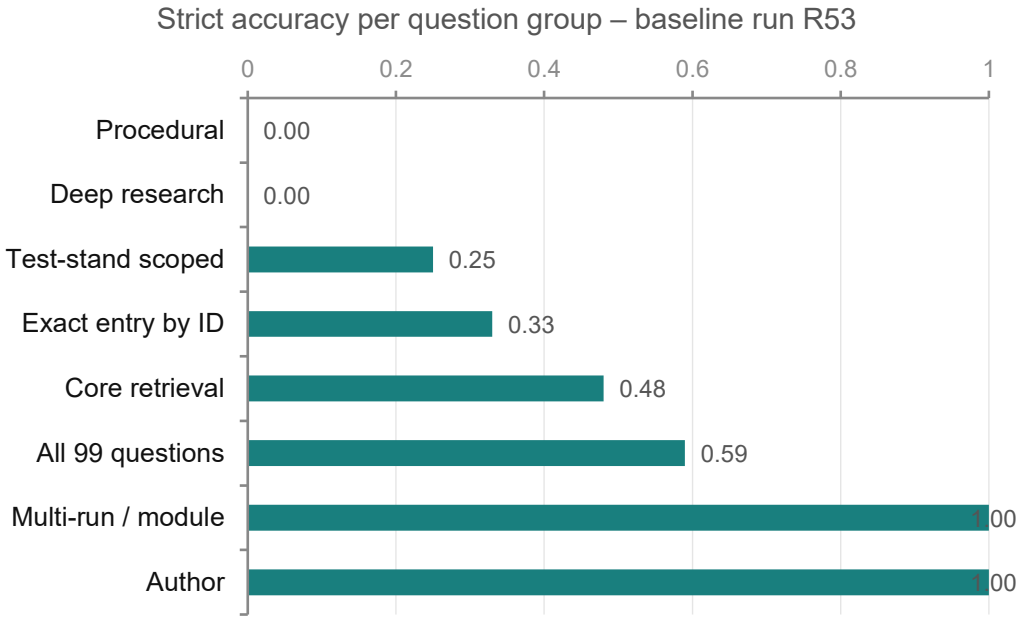
**229**

**cooldown runs**

Ramp rates, quench currents, dwell times, module and feedbox IDs, cross-references to earlier entries – per run, per module.

**Breadth vs. depth.** A shift log spans a whole machine in narrative form. A test-facility log spans one magnet in numbers. The questions people ask it are entity-scoped, numeric and temporal - "which Corrector Magnet of QDM-030 quenched above 230A in any cooldown run?" - and similarity search alone does not answer them.

# Where plain RAG breaks.




**Baseline: vector retrieval, one LLM pass - naive RAG pipeline.**

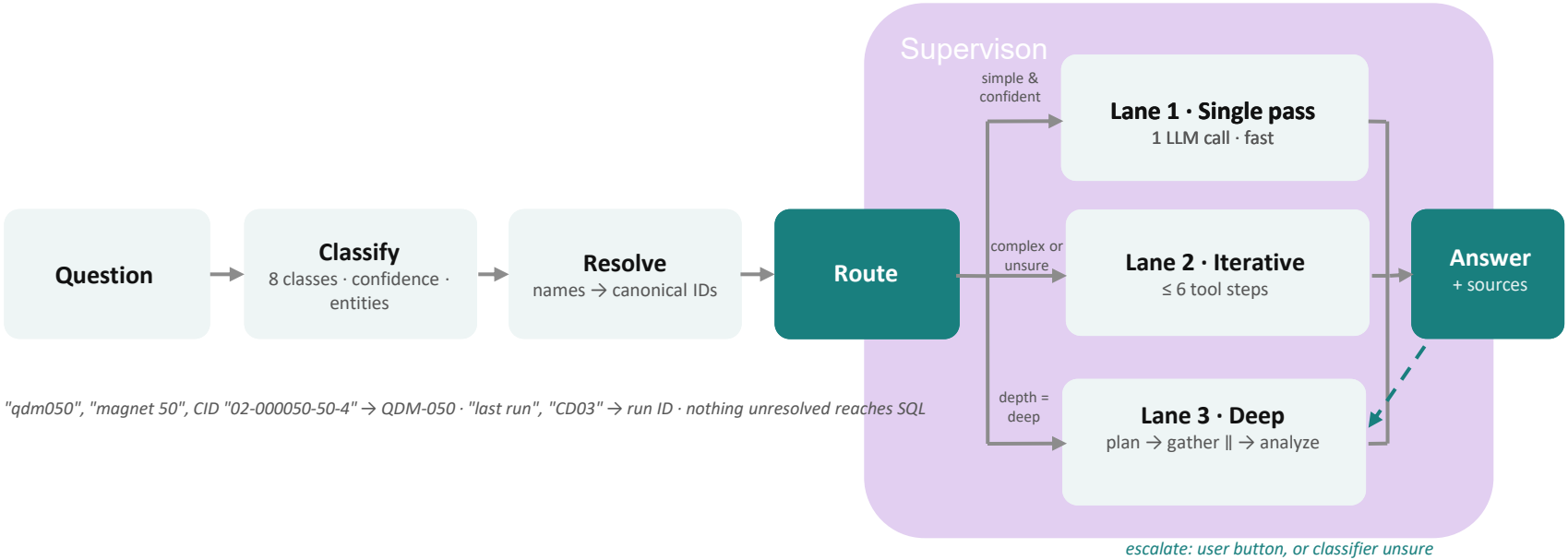
Similarity search finds look-alikes. It is fine for "what happened to module X" and fails outright on everything that needs structure or iteration:

- exact counts, totals, most recent entries
- procedures assembled from many entries
- analyses across several runs or modules
- scope limited to one test stand (FB1–FB4)

*Fixed question set (99 questions in 13 groups now), re-run after every change.*

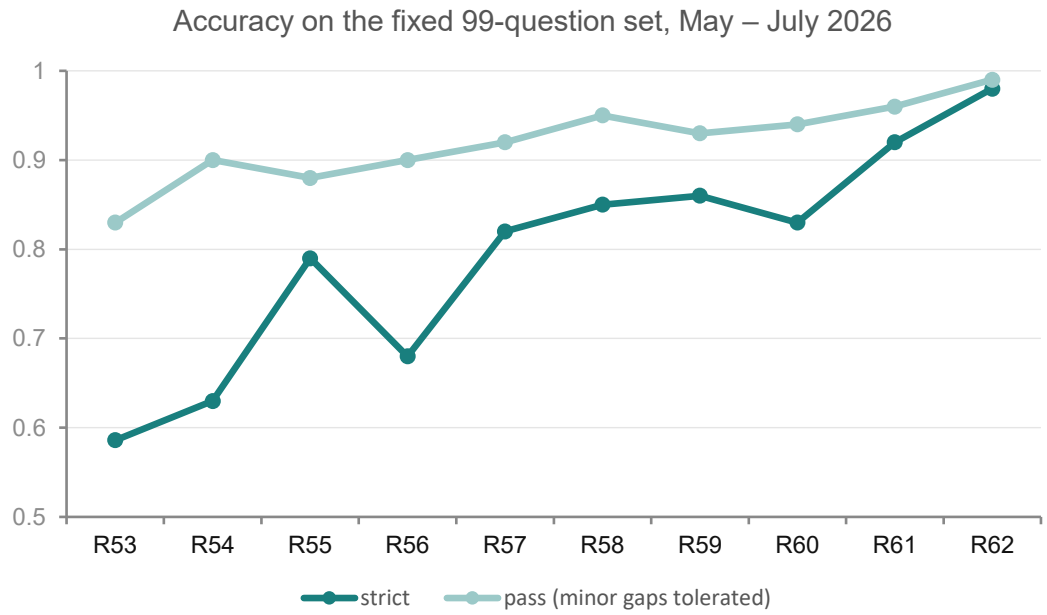
*Ground truth is not hand-labelled - each question carries a SQL recipe that re-derives it from the live database.*

# Classify, route, escalate.




*Bounded effort at every level: one call, six tool steps, or a plan you can read before it runs. Every answer cites entry IDs; ambiguous names produce a caveat, not a guess.*

# 0.59 to 0.98 in ten iterations.



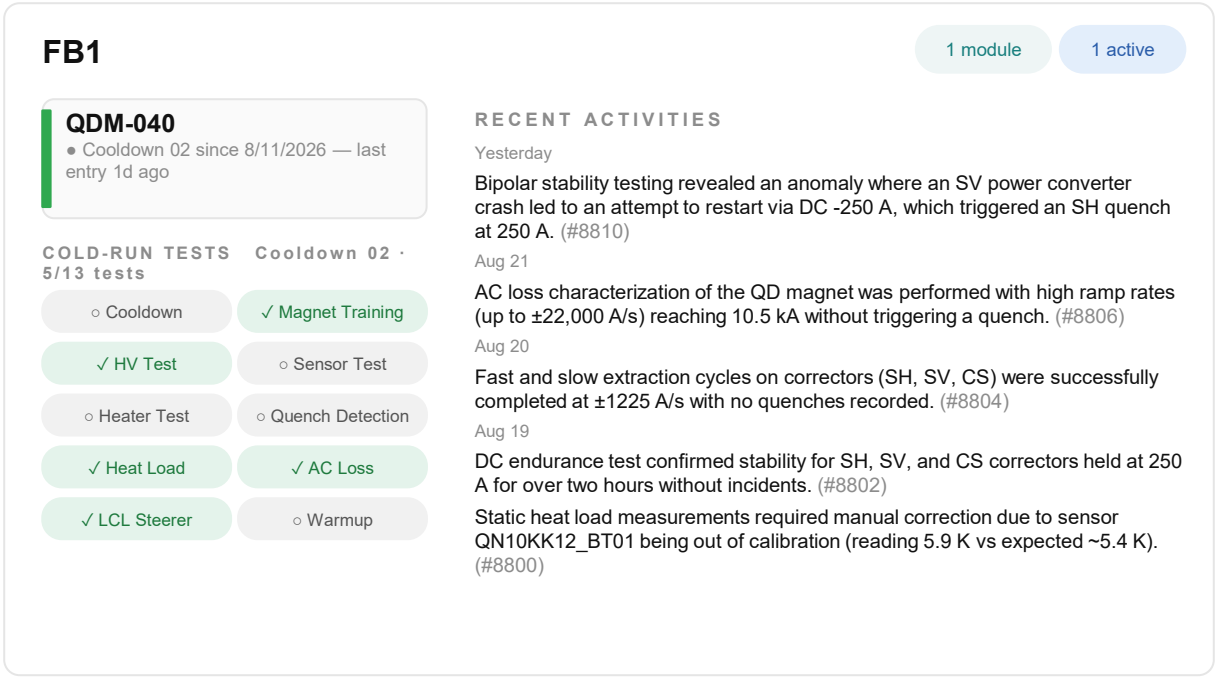

**0.98** **strict accuracy at R62**

**What moved the curve**

- procedural 0.00 → 1.00
- deep research 0.00 → 0.83
- test-stand scoped 0.25 → 1.00
- exact entry fetch 0.33 → 1.00
- core retrieval 0.48 → 0.96

*Temporal, multi-run and author questions were solid from the start and held. Shown: R53–R62 on the 99-question set;*

# It reads the logbook so nobody has to.




### Autonomous dashboard

Per-feedbox module state, cold-run test coverage and a nightly activity digest with entry IDs - no question needed. Each summary links to its sources; the state on the left is the live one from this morning.

### Ask-AI search

Plain-language questions, depth auto or deep, escalation offered on every answer. "Verify against the original" is built into the UI.

Live ingestion from the STF ELOG, containerised, GSI network.

GSI INNOVATIONFUND

# Production by year-end. Your logbook next.

## Where we are

- D1 RAG system with web UI – delivered
- D2 GSI integration: live ELOG, on-prem inference, dashboard - delivered
- D3 knowledge graph & evaluation - in evaluation, incl. user test against keyword-search baseline
- D4 production operation & documentation - Q4. Needs: an on-prem VM, LDAP auth, an ELOG service account

**Funded by the GSI Innovation Fund, Mar–Dec 2026**

**External industry pilot partner on board**

**Searching for further Partners – Industry and Research**

## Bring your logbook

**The pipeline is a template.** Any ELOG with structured entries - module, device, run, date - can be onboarded and evaluated with the same harness. Domain-specific parts: one fact schema, one set of relation rules.

**What it costs you**

- read to your ELOG Instance (service account)
- Meta Data Structure

**What you get**

- AI Assisted Search - find your Information in the “Needlestack” quickly

*>> Next: explicit ontology layer; same pattern on archived settings and observables.*

a.warth@gsi.de · +49 6159 71-3044

# Backup

# Recovery by question group (R53–R62).

| Group | R53 | R54 | R55 | R57 | R58 | R59 | R60 | R61 | R62 ▲ | Net |
|---|---|---|---|---|---|---|---|---|---|---|
| R 25 | 0.48 | 0.56 | 0.76 | 0.84 | 0.84 | 0.72 | 0.64 | 0.84 | 0.96 | ▲ +.48 |
| N 13 | 0.69 | 0.46 | 0.46 | 0.77 | 0.77 | 0.77 | 0.69 | 0.77 | 1.00 | ▲ +.31 |
| MD 8 | 0.63 | 0.75 | 1.00 | 1.00 | 1.00 | 0.88 | 0.88 | 0.88 | 1.00 | ▲ +.38 |
| TQ 7 | 1.00 | 0.86 | 0.71 | 0.71 | 0.86 | 1.00 | 1.00 | 1.00 | 1.00 | = ±0 |
| DE 6 | 0.83 | 0.83 | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | ▲ +.17 |
| MM 6 | 1.00 | 1.00 | 1.00 | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | = ±0 |
| EM 6 | 0.33 | 0.17 | 1.00 | 1.00 | 0.67 | 1.00 | 0.83 | 1.00 | 1.00 | ▲ +.67 |
| DR 6 | 0.00 | 0.50 | 0.83 | 0.67 | 0.33 | 0.67 | 0.83 | 1.00 | 0.83 | ▲ +.83 |
| AG 5 | 0.80 | 0.60 | 0.80 | 0.80 | 0.80 | 0.80 | 1.00 | 1.00 | 1.00 | ▲ +.20 |
| EC 5 | 0.60 | 0.60 | 0.80 | 0.80 | 1.00 | 1.00 | 0.80 | 1.00 | 1.00 | ▲ +.40 |
| TS 4 | 0.25 | 0.25 | 0.50 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | ▲ +.75 |
| AU 4 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | = perfect |
| PR 4 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | ▲ +1.00 |
| All 99 Q | 0.59 | 0.63 | 0.79 | 0.82 | 0.85 | 0.86 | 0.83 | 0.92 | 0.98 | ▲ +.39 |



The recovery is led by the groups that had failed outright at R53 — **procedural +1.00, deep-research +0.83, test-stand +0.75, exact-match +0.67** — and by **core-retrieval nearly doubling (+0.48)**. Three groups (temporal, multi-run, author) were already solid and held.

**13 groups, 99 questions**

R core retrieval · N numeric/recency · MD module diversity · TQ temporal · DE German-language · MM multi-run · EM exact entry · DR deep research · AG aggregate · EC edge cases · TS test-stand · AU author · PR procedural

Strict = complete answer, every number and ID correct. A single wrong count fails the question. Each record carries expected route, class, entities, source entry IDs and an oracle SQL recipe; a judge rubric scores the rest.

*The groups that failed outright at baseline are the ones that need structure and iteration – exactly where the graph and the agent lanes act.*

# Hybrid retrieval, one query.

**QUERY "What is currently done on QDM-040?"**

**KNOWLEDGE GRAPH**
Structural traversal
Exact counts, relations, current module state

*Derived from entry metadata – no LLM*

**VECTOR SEARCH**
Semantic similarity
Embed → cosine → top-k

*Finds cases described differently*

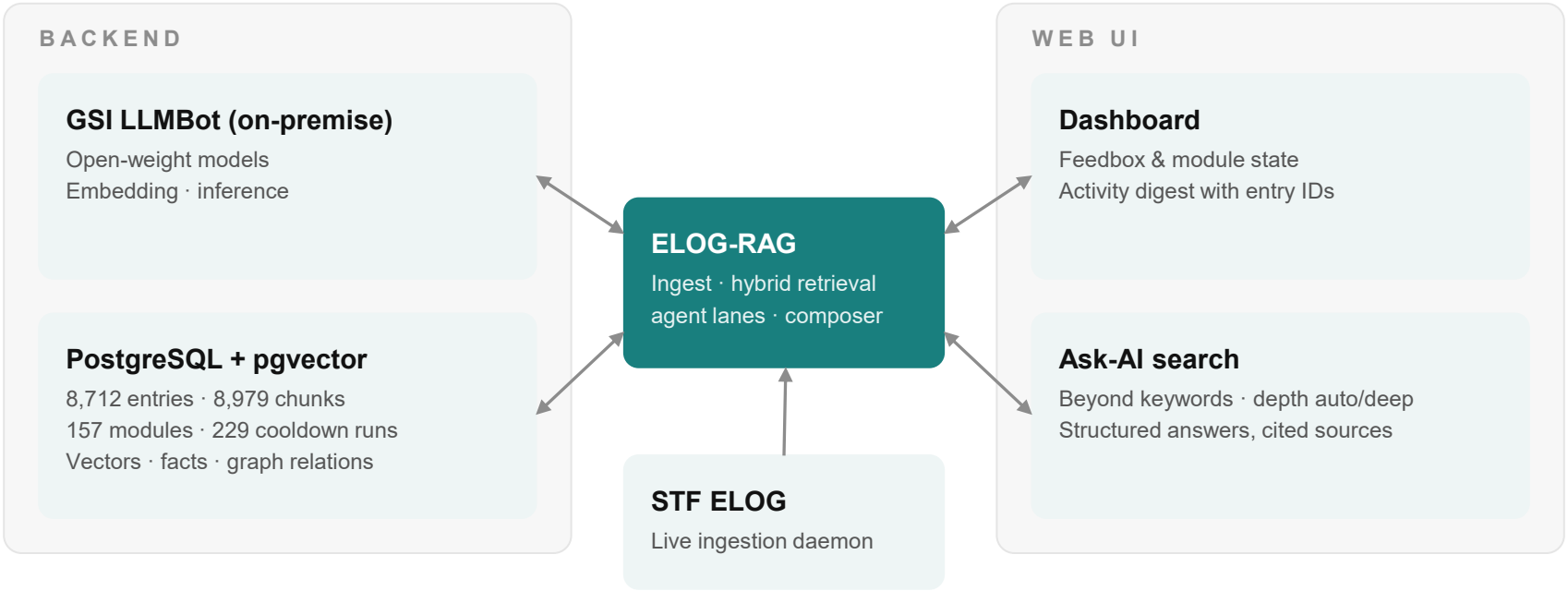
**KEYWORD SEARCH**
BM25 full-text
"QDM-040", IDs, part numbers

*Exact identifiers never get lost*

**FUSION & RERANKING deduplicate · merge · rerank · top-k → LLM synthesis with citations**

ANSWER + SOURCES #8746 · #8747 · … – structured, semantic and lexical signals fused into one ranked answer.

# Topology.


BACKEND
GSI LLMBot (on-premise)
Open-weight models
Embedding · inference
PostgreSQL + pgvector
8,712 entries · 8,979 chunks
157 modules · 229 cooldown runs
Vectors · facts · graph relations
ELOG-RAG
Ingest · hybrid retrieval
agent lanes · composer
STF ELOG
Live ingestion daemon
WEB UI
Dashboard
Feedbox & module state
Activity digest with entry IDs
Ask-AI search
Beyond keywords · depth auto/deep
Structured answers, cited sources