

Trust the Query, Not the Answer

Auditable Text-to-SPARQL over the GLEIF Register

Jannic Alexander Cutura

European Central Bank, Sonnemannstrasse 20, 60314 Frankfurt am Main, Germany

DSTI School of Engineering, 4 Rue de la Collégiale, 75005 Paris, France

jannic.cutura@dsti.institute

Abstract

Ask which subsidiaries a conglomerate owns and a language model returns a plausible list, unverifiable and partly wrong; in compliance or supervision that will not do. So instead of trusting the answer, a local open-weight model translates each question into an auditable SPARQL query over the GLEIF register, returning the query and its source records. A 20B model answers 86.7% of held-out questions correctly, always with a valid query and declining the unanswerable; a five-times larger open model and a closed frontier model reach 87.8 and 90.0%, neither gap distinguishable from noise on 90 questions. The construction fits any ontology-described source, though we test one; code and benchmark are released.

1 Introduction

In domains such as anti-money-laundering screening, credit-risk assessment and banking supervision, an answer is useful only if it can be audited, that is, traced to a source record and confirmed. A large language model queried directly does not meet this requirement. Asked about corporate structure or ownership it returns fluent prose whose correctness cannot be established without redoing the work by hand, and whose mistakes are phrased with the same confidence as its correct answers; in a regulatory setting such a plausible error is worse than an explicit refusal, because it will be acted upon.

Auditable question answering over reference data needs two things. The first is a reliable structured source that records the objects of a domain and the relationships among them under stable identifiers; a knowledge graph is the natural form, and many regulated domains now publish one. We

demonstrate on the GLEIF register of legal entities (§2), a concrete public example in which nothing is specific to corporate structure. The second is a way to query that source in natural language, because SPARQL is expressive but unfamiliar to the analysts who would use it. Mapping a plain-English question to an executable query supplies both at once: the user writes no SPARQL, while the query, its result rows and the underlying records stay open to inspection. Answering enterprise questions through such a graph-and-SPARQL representation is also more accurate than answering directly over the raw schema with the same model (Sequeda et al., 2024).

We give a deliberately minimal instance of this recipe (Figure 1). A local open-weight model translates questions into SPARQL over GLEIF, a validation guard checks each query before execution and a bounded repair loop corrects it, and every answer is returned with its query and cited records. GLEIF is a single, comparatively simple graph and the questions we pose are routine ones, but the construction, a model writing queries against any source described by an ontology, is general. The parts are standard: schema-conditioned generation, a validity check before execution, a bounded repair. What is new is their integration over a real regulatory register and the benchmark that measures it. The contribution is thus a concrete, reproducible demonstration: an execution-verified benchmark of 144 questions, split into a development and a held-out set, measuring how well a local open model writes correct, auditable queries, including out-of-scope questions it must decline; a local, privacy-preserving text-to-SPARQL system over GLEIF whose every answer is grounded in its query and cited records; and released code, prompts and benchmark, reproducible against a pinned snapshot.

The views expressed in this paper are those of the author and do not necessarily reflect those of the European Central Bank or the Eurosystem. Code, prompts and benchmark: <https://github.com/JannicCutura/finnlp-2026>.

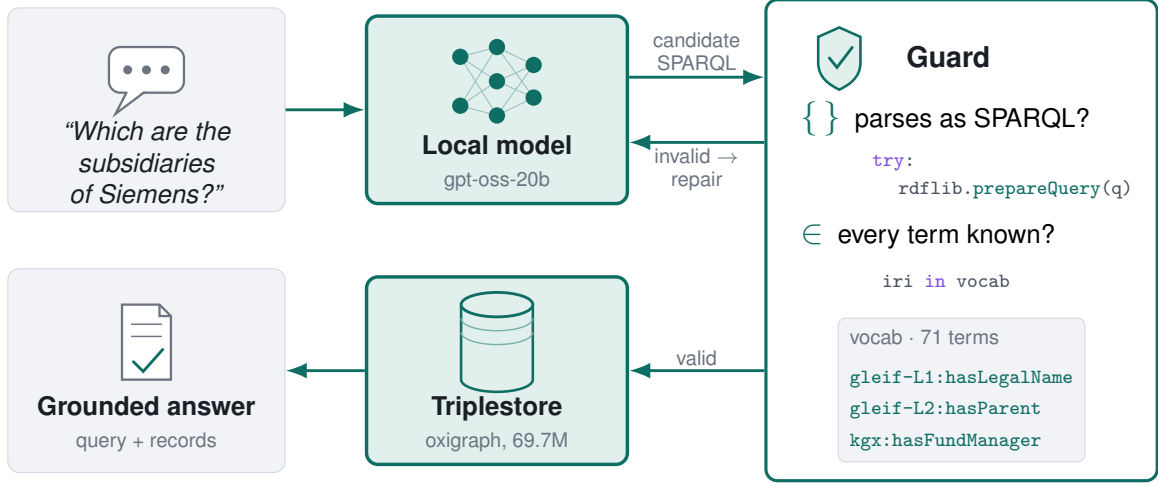


Figure 1: **The querying pipeline.** A local model writes candidate SPARQL; a guard checks that it parses and that every term exists in the graph vocabulary, returning its errors for a bounded repair; a valid query executes and the answer is returned with the query and the source records.

2 Background: the GLEIF Graph

GLEIF publishes the LEI register as CC0 open data in two levels. Level 1 records each entity’s legal name, jurisdiction, ISO 20275 legal form, category and status; Level 2 records relationships between entities that hold LEIs (accounting consolidation, international branches and fund structures) under an official RDF ontology. Our snapshot (20260723–1600) holds 3,382,301 Level 1 records (1,940,941 ISSUED) and 481,557 Level 2 relationships. The gain over asking the model directly is concrete: asked which entities are subsidiaries of Siemens, a general-purpose chat assistant returns about a dozen plausible names, none carrying the identifier needed to verify them, whereas the register returns exactly 452 entities that report Siemens Aktiengesellschaft as their direct consolidation parent, each traceable to its source record by LEI.

No prior academic work puts an LLM text-to-SPARQL layer over GLEIF: existing uses are analytic and hand-written (Lüdemann et al., 2020; Trypuz et al., 2016), GLEIF’s own assistant and an MCP wrapper disclose no query-generation mechanism and publish no evaluation (Global Legal Entity Identifier Foundation, 2026; Safonova, 2026), and text-to-SPARQL research targets encyclopaedic or synthetic graphs (Kosten et al., 2023; Marx et al., 2025) rather than a real regulatory register.

3 System

The system runs locally (Figure 1). We serve gpt-oss-20b, a 20B open-weight model, through LM Studio at temperature 0 with prompting only; open-weight models are competitive code generators (Hui et al., 2024) and have been used for SPARQL generation (Brei et al., 2025). We generate each query in a single shot rather than through the multi-turn graph exploration of agentic KGQA systems (Liu et al., 2024; Brei et al., 2025; Perevalov and Both, 2025); the guard, not iterative search, is what keeps the output valid. Each question is wrapped with a schema of the graph¹ and sent to the model, which returns candidate SPARQL; a pre-execution guard then checks that the query parses and that every term it uses exists in the 71-term vocabulary, which is all 27 predicates the graph populates plus the 44 classes and code-list values they take: a cheap defence against hallucinated predicates in the ontology-check-and-repair line of Allemang and Sequeda (2024). On a syntax, vocabulary or runtime error the guard returns the fault with a closest-match suggestion for a bounded repair of at most two retries, but never retries on an empty result, which is the correct answer to a question the register cannot answer and, at execution time, is indistinguishable from

¹The GLEIF Golden Copy is converted to RDF and loaded into oxigraph, a disk-backed SPARQL 1.1 store of 69.7M triples pinned to 20260723–1600. The shipped ontology omits classes for GLEIF’s fund relationships and its reified records make multi-hop traversal slow, so we add documented shortcut predicates and advertise them in the schema.

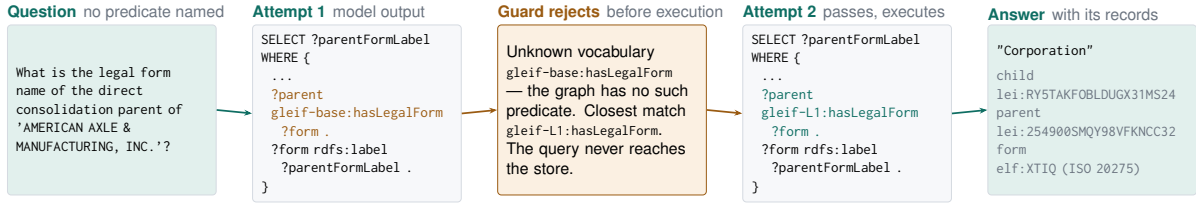


Figure 2: **One held-out question, end to end**, as the run log recorded it (item h55; triples common to both attempts elided as . . .). The guard rejects a predicate the graph does not have and returns the closest match; the model repairs on the first retry. This is the whole audit trail an analyst is shown.

a wrong query that returns nothing. Every answer carries its query, rows and GLEIF identifiers, so it is never separable from the evidence for it; the closest prior design, retrieval-plus-validation over a real domain graph (Emonet et al., 2024), is neither local nor prompt-only. Figure 2 traces one held-out question through the whole path, and is also what the user is shown. What the guard certifies is narrower than *validation* suggests: that a query parses and names only terms the graph has, never that it means what was asked. A query can pass and still be wrong, which is not a gap in the design but the point of it, since the error then arrives as a readable query returning the wrong rows rather than as fluent prose.

4 Evaluation and Results

Question types were fixed in a specification before instantiation, around a KYC analyst, a prudential supervisor and a fund-operations user, then written in business language that names no predicate or IRI. Six tiers span lookup, single-hop, multi-hop, attribute filter, aggregate and unanswerable.² Each question carries a reference query whose execution result is the gold answer, and every reference query is executed against the pinned snapshot before use, so a gold answer cannot silently drift from the graph. The 54-question development set informed three schema revisions; the 90-question held-out set was authored afterwards against a frozen prompt and schema and informed none, so it estimates generalisation rather than serving as a tuning target. No question in either set was written with sight of any model output.

Our metric is execution accuracy under a relaxed

²The unanswerable tier holds questions the register cannot answer, such as an ownership percentage, which does not exist because GLEIF Level 2 records accounting consolidation rather than equity ownership, or a fact about an entity absent from the graph. It is capped at about a fifth of the set so that indiscriminate declining cannot score.

Tier	20B		120B
	Dev	Held	Held
a lookup	10/10	10/15	12/15
b single-hop	10/10	14/14	14/14
c multi-hop	8/8	13/15	14/15
d attribute filter	7/7	14/15	14/15
e aggregate	6/7	13/15	11/15
f unanswerable	12/12	14/16	14/16
Execution accuracy	98.1%	86.7%	87.8%
95% CI	90–100	78–92	79–93
QALD macro-F1	0.981	0.897	0.893
Validity rate	100%	100%	98.9%
First-try / repaired	48/6	82/8	80/9

Table 1: Per-tier exact match and summary metrics, with Wilson 95% intervals on accuracy. gpt-oss-20b on the 54-question development and 90-question held-out sets, and the five-times larger gpt-oss-120b on the held-out set; the one 120B item that induced a generation loop is counted wrong. Temperature 0, repair budget 2. The two held-out intervals overlap almost completely.

exact match (Pan et al., 2025) that compares result rows as a set, ignoring column names and row order, reported with a Wilson 95% confidence interval, alongside QALD macro-F1 under the empty-answer convention (Usbeck et al., 2024) and a TrustSQL-style reliability score that rewards a correct answer or a correct decline and penalises a confident wrong answer by a factor c (Lee et al., 2024a,b). Absolute accuracy is not comparable across studies: these figures run well above the 45 to 55% of general benchmarks (Sequeda et al., 2024; Kosten et al., 2023; Li et al., 2023) because GLEIF is one register with a thin schema, not because the queries are better written.

The system reaches 98.1% on the development set and 86.7% on the held-out set (Table 1), every query valid and none unresolved, so no answer is returned without a query behind it. Development accuracy rose from 92.6% to 98.1% after the three schema revisions diagnosed from that set, and should therefore be read as a tuned ceiling, not an

estimate of generalisation. The ablation (Table 2) locates each component’s contribution. The raw model already writes a valid query 91.1% of the time and answers 80.0% correctly; the bounded repair loop, fed the guard’s diagnostics, lifts these to 100% and 86.7%, every repair succeeding on the first retry. Both conditions come from the same run on the same questions, so the gain is paired: exact McNemar gives $p = 0.031$, and the discordant counts say more than the p does, being six and zero. The loop fixed six questions and broke none. Removing the in-prompt vocabulary drops accuracy to 32.2%, and removing the schema entirely collapses it to 13.3%, the ontology-mediation effect (Allemang and Sequeda, 2024) measured on this graph.

Where the errors do *not* occur is the point: the model declines every out-of-scope and absent-entity question. The two unanswerable-tier misses are recorded-exception questions, which the register can answer by returning the exception and which the model returned in the wrong shape, not failures to decline. Recognising true unanswerability is thus the easy part, and reliability stays positive even under a heavy penalty for a confident wrong answer ($c=1, 3, 5$ give $+0.74, +0.50, +0.26$), because only 11 of 90 answers are confident and wrong. All twelve held-out errors are queries that executed and returned the wrong rows, never invalid or hallucinated ones; they fall into string matching, literal typing, wrong predicate or scope, and result shape (Appendix A). Each is visible in the query itself, the property the setting demanded. The weakest held-out tier is lookup (10/15), where the misses are string matching, case and the choice of match operator; a case-insensitive

Held-out (20B)	Valid	Acc
Full system (guard+repair)	100%	86.7%
– guard & repair	91.1%	80.0%
– vocabulary	100%	32.2%
– schema	100%	13.3%

Table 2: Ablation on the held-out set (gpt-oss-20b). Top: the full system against the raw model, with the guard and repair loop removed. Bottom: the full system with the 71-term vocabulary, then the whole schema, stripped from the prompt. The guard and repair loop raise accuracy from 80.0 to 86.7% and validity to 100%; paired over the same questions, the gain is six items fixed and none broken (exact McNemar, $p = 0.031$). The in-prompt vocabulary accounts for most of the rest.

Held-out	Acc	95% CI	Valid
<i>Local, open-weight</i>			
gpt-oss-20b	86.7%	78–92	100%
gpt-oss-120b	87.8%	79–93	98.9%
Qwen2.5-Coder-32B	85.6%	77–91	97%
Llama-3.3-70B	70.0%	60–78	97%
<i>Non-local reference (not deployable)</i>			
gpt-5	90.0%	82–95	100%

Table 3: Held-out execution accuracy across model families (90 questions). The gpt-oss and Qwen models cluster within one another’s 95% intervals; Llama-3.3-70b is a clear outlier below them, and the closed frontier model leads. gpt-5 is a non-local reference, not deployable where the data cannot leave the machine; it denotes the OpenAI gpt-5 API model, accessed 29 July 2026.

matching instruction is an obvious fix we did not apply, since it would have meant unfreezing the prompt. Queries run in well under a second except full-population aggregates (~ 14 – 28 s); latency is dominated by generation (~ 15 – 40 s), not the store. Model size is not the bottleneck, nor is open versus closed; training focus is (Table 3). The gpt-oss models and Qwen2.5-Coder-32B cluster at 85.6–87.8%, within one another’s intervals, and gpt-5, an API-only reference that cannot be deployed where the data must stay on the machine, leads by about three points at 90.0%. On the same paired test none of those gaps against the 20B separates from noise ($p \geq 0.38$); the one model that does is Llama-3.3-70b, the largest open model we test and the weakest at 70.0% (-16.7 points, $p = 0.0007$), which fails mostly on single- and multi-hop joins: a general instruction model gives up ground to code-tuned ones on structured generation, which is what the task rewards. gpt-5 also produced three queries that were valid yet computationally runaway and caught by timeout, a further argument for a bounded budget.

5 Conclusion

A local 20B model, given the schema, a guard and a two-retry repair loop, writes auditable SPARQL over GLEIF for 98.1% of a development set and 86.7% of a harder 90-question held-out set, every answer traceable to its query and records and every out-of-scope question declined. The point is not the accuracy but that each answer, right or wrong, sits in a query an analyst can read and re-run against a pinned snapshot; nothing in the construction is specific to corporate structure, though we test one.

Limitations

The evaluation is small (144 questions), which the reported confidence intervals reflect, and uses a single machine. The GLEIF graph is thin (27 populated predicates), so absolute accuracy is not comparable to open-domain KGQA. The development number follows three schema revisions diagnosed from that set, so only the held-out number estimates generalisation. Level 2 is accounting consolidation, not ownership, and the snapshot is a single 8-hourly publish (20260723-1600), so gold answers hold only for that pin. We compare open-weight models from different families and one closed frontier model as a non-local reference, at temperature 0 and all on public GLEIF data; a run-to-run variance study remains future work. Beyond the execution gate above, a single author wrote and checked both sets: no second annotator scored them and we report no agreement figure. The personas shaped the question specification and nothing beyond it, so the questions are ours rather than an analyst's, and no user was observed working with the system. We test one graph, so the claim that the construction carries to any ontology-described source is an argument from its shape, not a measurement. A single graph is also the easy case: real data sits across relational databases, graph and network stores, document stores, and text or PDF files, each with its own query language. Extending the same generate, guard and repair loop across these sources, routing each question to the store that can answer it and keeping every answer decomposable into a checkable query, is the direction we intend to pursue.

Ethics Statement

GLEIF LEI data is CC0 public-domain reference data about legal entities, not natural persons. The register never records the identity of a natural person; a NATURAL_PERSONS reporting exception states only *that* a parent is a natural person, and never who that person is. The released evaluation slice contains only legal-entity reference data and our own questions and queries.

Disclaimer

The views expressed in this paper are those of the author and do not necessarily reflect those of the European Central Bank or the Eurosystem.

References

- Dean Allemang and Juan Sequeda. 2024. [Increasing the accuracy of LLM question-answering systems with ontologies](#). In *The Semantic Web – ISWC 2024*, volume 15233 of *Lecture Notes in Computer Science*, pages 324–339. Springer.
- Felix Brei, Lorenz Bühmann, Johannes Frey, Daniel Gerber, Lars-Peter Meyer, Claus Stadler, and Kirill Bulert. 2025. [ARUQULA – an LLM based Text2SPARQL approach using ReAct and knowledge graph exploration utilities](#). In *Proceedings of the First International TEXT2SPARQL Challenge, co-located with Text2KG at ESWC 2025*, volume 4094 of *CEUR Workshop Proceedings*, pages 49–63. CEUR-WS.org.
- Vincent Emonet, Jerven Bolleman, Séverine Duvaud, Tarcisio Mendes de Farias, and Ana Claudia Sima. 2024. [LLM-based SPARQL query generation from natural language over federated knowledge graphs](#). In *Proceedings of SWAT4HCLS 2024 (CEUR-WS)*, volume 3953 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Global Legal Entity Identifier Foundation. 2026. GLEIF AI. <https://ai.gleif.org/>. Production natural-language assistant over LEI data; architecture undisclosed. Accessed 2026.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. [Qwen2.5-Coder technical report](#). *arXiv preprint arXiv:2409.12186*.
- Catherine Kosten, Philippe Cudré-Mauroux, and Kurt Stockinger. 2023. [Spider4SPARQL: A complex benchmark for evaluating knowledge graph question answering systems](#). In *2023 IEEE International Conference on Big Data (BigData)*. IEEE.
- Gyubok Lee, Woosog Chay, Seonhee Cho, and Edward Choi. 2024a. [TrustSQL: Benchmarking text-to-SQL reliability with penalty-based scoring](#). *arXiv preprint arXiv:2403.15879*.
- Gyubok Lee, Sunjun Kweon, Seongsu Bae, and Edward Choi. 2024b. [Overview of the EHRSQL 2024 shared task on reliable text-to-SQL modeling on electronic health records](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 644–654. Association for Computational Linguistics.
- Jinyang Li, Binyuan Hui, Ge Qu, Jiayi Yang, Binhua Li, Bowen Li, et al. 2023. [Can LLM already serve as a database interface? a BIG bench for large-scale database grounded text-to-SQLs](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*.
- Shicheng Liu, Sina J. Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, and Monica S. Lam. 2024. [SPINACH: SPARQL-based information navigation for challenging real-world questions](#). In *Findings of the Association for Computational Linguistics*:

- EMNLP 2024, pages 15977–16001. Association for Computational Linguistics.
- Niklas Lüdemann, Ageda Shiba, Nikolaos Thymianis, Nicolas Heist, Christopher Ludwig, and Heiko Paulheim. 2020. [A knowledge graph for assessing aggressive tax planning strategies](#). In *The Semantic Web – ISWC 2020*.
- Edgard Marx, Paulo Viviurka do Carmo, Marcos Gôlo, and Sebastian Tramp. 2025. [Preface of the first international TEXT2SPARQL challenge \(TEXT2SPARQL’25\)](#). In *Proceedings of the First International TEXT2SPARQL Challenge, co-located with Text2KG at ESWC 2025*, volume 4094 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Xueli Pan, Victor de Boer, and Jacco van Ossenbruggen. 2025. [FIRESPARQL: A LLM-based framework for SPARQL query generation over scholarly knowledge graphs](#). *arXiv preprint arXiv:2508.10467*.
- Aleksandr Perevalov and Andreas Both. 2025. [Text-to-SPARQL goes beyond English: Multilingual question answering over knowledge graphs through human-inspired reasoning](#). In *Proceedings of the First International TEXT2SPARQL Challenge, co-located with Text2KG at ESWC 2025*, volume 4094 of *CEUR Workshop Proceedings*, pages 77–93. CEUR-WS.org.
- Olga Safonova. 2026. [gleif-mcp-server: MCP server for GLEIF LEI database access](#). <https://github.com/olgasafonova/gleif-mcp-server>.
- Juan Sequeda, Dean Allemang, and Bryon Jacob. 2024. [A benchmark to understand the role of knowledge graphs on large language model’s accuracy for question answering on enterprise SQL databases](#). In *Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*.
- Robert Trypuz, Dominik Kuziński, and Mirosław Sopek. 2016. [General legal entity identifier ontology](#). In *Joint Proceedings of the Posters and Demos Track of ISWC 2016 (CEUR-WS)*, volume 1660 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Ricardo Usbeck, Xi Yan, Aleksandr Perevalov, Longquan Jiang, Julius Schulz, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, and Andreas Both. 2024. [QALD-10 – the 10th challenge on question answering over linked data](#). *Semantic Web*, 15(6).

A Held-out error taxonomy

Error type	n
String / entity matching	4
Literal typing	2
Wrong predicate / value / scope	3
Result shape / aggregation	3
Total	12

Table 4: The twelve held-out errors, hand-classified. Every one is a query that executed and returned the wrong rows; none is invalid or hallucinated, since the guard rejects those, so the fault is always visible in the query. String matching: case or match operator (h01, h05, h34, h38). Literal typing: `xsd:date` against the graph’s `xsd:dateTime` (h22, h75). Wrong predicate, value or scope: transitive for direct, a wrong ELF code, an over-cautious decline (h15, h62, h68). Result shape: COUNT or DISTINCT multiplicity, or the wrong columns (h03, h28, h82).