Design and Implementation of Aletheia: an industrial semantic layer for generative artificial intelligence agents on legacy systems

Ignacio García Marín

Universidad Pontificia Comillas ICAI, Madrid, Spain

ARTICLE INFO

Article type: Master's Thesis paper

Programme: MII–MIINT

Academic year: 2025/26

Keywords

Generative AI; industrial metadata; legacy systems; semantic layer; RAG; knowledge graph; human-in-the-loop; Industry 4.0

ABSTRACT

Industrial organizations seek to exploit generative artificial intelligence over operational data, yet legacy relational systems remain semantically opaque, fragmented and difficult to govern. This paper presents Aletheia, a metadata-first semantic layer that transforms heterogeneous industrial schemas into validated, traceable knowledge for AI agents. The architecture combines technical profiling, LLM-based enrichment, embedding-assisted relationship discovery, deterministic quality gates, RAG artefacts and graph topology construction. AXON complements the pipeline by converting downstream failures into human-reviewed semantic corrections. Evaluation in a real steel-manufacturing environment covered 3 databases, 2,118 tables and 31,735 fields, producing 4,688 Markdown chunks and 12.48 million indexable characters. Downstream evaluation achieved 97.29% relevance, 88.72% correctness, 88.95% faithfulness and 81.34% transparency. AXON obtained 24/24 deterministic PASS results, while its LLM judge scored problem resolution, technical correctness and actionability at 4.27, 4.18 and 4.32 out of 5. The results support governed semantic infrastructure as a prerequisite for reliable industrial generative AI.

1. Introduction

Heavy industry has accumulated decades of operational information across relational databases, manufacturing execution systems, mainframes and specialised analytical repositories. Although these systems remain operationally robust, their meaning is frequently encoded in abbreviated names, undocumented composite keys, implicit relationships and tacit expert knowledge. The resulting semantic debt creates a practical paradox: data are physically available, but their reliable analytical use remains dependent on a small number of specialists who understand the internal topology of the information landscape.

Transformer-based large language models (LLMs) create an opportunity to democratise access to industrial information through conversational interfaces and automated analysis [2]. However, directly connecting a generative agent to an opaque catalogue does not solve the underlying problem. A syntactically valid SQL query may still join unrelated entities, ignore grain mismatches, invent fields or reproduce a misleading interpretation of a process. These failure modes are consistent with the broader grounding and hallucination risks identified in retrieval-augmented systems [3]. In operational settings, they may distort decisions concerning quality, cost, energy, maintenance or production performance.

This work addresses the infrastructure that must exist before autonomous analytical agents can operate safely. Rather than treating retrieval-augmented generation (RAG) as a document search problem, the proposed approach models industrial metadata as governed knowledge. The central research question is therefore: how can heterogeneous legacy schemas be transformed into a semantic, auditable and continuously maintained representation that generative AI agents can consume without initially replicating raw production data?

The paper makes four contributions. First, it defines a metadata-first architecture that combines probabilistic semantic inference with verifiable structural controls. Second, it generates two complementary representations: atomic Markdown artefacts for retrieval and a graph topology for relational reasoning. Third, it introduces AXON, a human-supervised feedback mechanism that converts downstream failures into controlled semantic patches. Fourth, it evaluates the complete ecosystem at industrial scale and discusses its methodological and operational limitations.

2. Background and related work

2.1 Retrieval-augmented generation and semantic debt

RAG augments a generative model with external evidence retrieved at inference time, reducing dependence on parametric memory and improving traceability [1]. Surveys of the field show, however, that retrieval quality, evidence construction and evaluation remain separate engineering challenges rather than guarantees of correctness [3]. Conventional implementations typically divide documents into chunks, embed them in a vector space and retrieve the nearest passages to a user query. This paradigm is effective when relevant knowledge is primarily textual, but industrial database reasoning is inherently relational. Correct analysis depends not only on locating a table description, but also on understanding valid join paths, composite keys, cardinalities, grain, system boundaries and constraints that prevent fan-out or double counting.

Metadata enrichment with LLMs can accelerate documentation by inferring descriptions, business concepts and likely relationships from schema names, data types and bounded samples. Aletheia performs this inference through Amazon Bedrock [13]. Nevertheless, model outputs remain probabilistic and can propagate plausible but incorrect assumptions. The challenge is therefore not simply to generate richer descriptions, but to embed inference in an engineering workflow that preserves provenance, validates

structure and applies safeguards against common LLM application risks [16].

2.2 Knowledge graphs and governed industrial AI

Knowledge graphs provide an explicit representation of entities and relationships and are particularly useful for multi-hop reasoning [5]. In the proposed system, graph topology complements vector retrieval rather than replacing it: Markdown chunks provide local semantic detail, while the graph constrains connectivity across the catalogue. This hybrid design is consistent with GraphRAG approaches that combine textual evidence with graph-based navigation [4].

The industrial context also imposes security requirements. Aletheia follows a metadata-first strategy: schemas, statistics, types, frequent values and controlled samples are processed, while complete transactional records and historical series remain inside the operational perimeter. This reduces unnecessary exposure and aligns the initial deployment with Zero Trust and industrial defence-in-depth principles [7], [8].

3. Problem definition and design requirements

The target environment contains heterogeneous SQL Server, analytical and legacy AS/400 sources whose schemas have evolved over decades. Referential integrity is often not explicitly declared, naming conventions are inconsistent and domain knowledge is distributed across stored procedures, historical queries and individual experience. A useful semantic layer must satisfy the following requirements:

- Scalability: process thousands of tables and tens of thousands of fields without relying on manual documentation.
- Traceability: preserve the origin of every generated description, relationship and correction.
- Structural reliability: prevent semantically plausible but physically invalid relationships from entering the governed catalogue.
- Security: avoid unnecessary replication of raw industrial data during the initial deployment stage.
- Dual consumption: support both semantic retrieval and relational path reasoning.
- Continuous maintenance: incorporate validated user feedback without allowing autonomous uncontrolled modifications.

These requirements led to a closed-loop architecture with three distinct roles. Aletheia constructs the semantic layer; Ariadne consumes it as a downstream conversational and analytical agent; AXON maintains it by diagnosing failures and proposing reviewed corrections. The separation prevents the consuming agent from silently rewriting the knowledge on which its own answers depend.

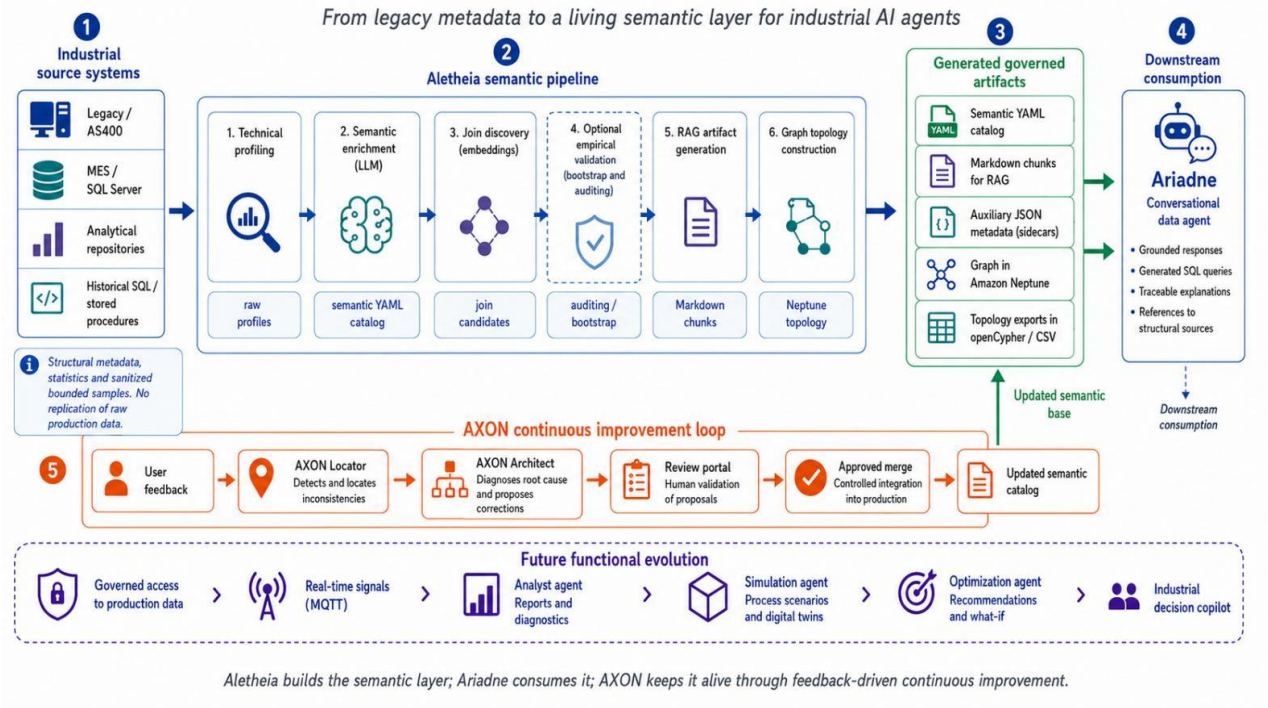

Fig. 1. Global Aletheia–AXON architecture. Aletheia builds the semantic layer, Ariadne consumes it, and AXON maintains it through feedback-driven review and controlled merge.

3.1 Research scope and success criteria

The scope deliberately excludes autonomous execution of production queries and direct control of industrial processes. The project focuses on the semantic and governance layer required before those capabilities can be considered. Success is therefore not defined by a single chatbot benchmark, but by the ability to generate a broad, structurally consistent and maintainable knowledge base that improves downstream performance while preserving human control.

Three complementary success criteria were established. The first is technical coverage: the pipeline must process heterogeneous sources at the scale of the real catalogue and generate valid artefacts without manual editing of every table. The second is downstream utility: an independent consuming agent must be able to retrieve the generated knowledge and produce relevant, correct and traceable outputs. The third is operational sustainability: failures observed after deployment must be convertible into bounded corrections with explicit review and low incremental cost.

This definition avoids a common evaluation error in generative systems: treating fluent output as evidence that the underlying knowledge is correct. In the proposed framework, fluency is secondary to provenance, admissibility and controlled change. Aletheia is considered successful when it constrains what an agent can reasonably claim, exposes the evidence behind that claim and provides a process for correcting the semantic substrate when real use reveals a defect.

4. System architecture and methodology

4.1 Technical profiling

The first stage creates normalised technical profiles from source systems. The profiler extracts schema identifiers, table and column names, data types, null ratios, cardinalities, representative value distributions and bounded sanitised samples. Profiles form a stable contract between on-premise extraction and cloud processing. Because the input is explicit and versioned, later semantic transformations can be reproduced and audited.

4.2 Semantic enrichment

Profiles are enriched using foundation models through Amazon Bedrock [13]. Prompt contracts request concise table purposes, column meanings, business concepts, warnings and confidence indicators. Outputs are stored as YAML rather than free text, enabling deterministic parsing, schema validation and version control. Defensive post-processing rejects malformed outputs, unknown columns and unsupported additions. The result is a semantic catalogue that remains linked to the physical schema from which it was derived.

4.3 Relationship discovery and validation

Candidate relationships are generated using a combination of lexical evidence, sentence embeddings, data-type compatibility, naming patterns, domain rules and LLM-assisted assessment. Sentence-BERT provides the methodological basis for efficient semantic comparison of short textual descriptions [6]. The objective is to recover implicit joins that are absent from formal database constraints. Candidate ranking reduces the search space and provides explanations for each proposed edge.

During development, an empirical validation module executed non-blocking existence checks and type controls against local sources. It was used to evaluate candidate joins and calibrate discovery thresholds. Because its marginal contribution did not justify the recurrent operational coupling and cost, it was removed from the normal automated production flow and retained as an exceptional audit and bootstrap mechanism. This distinction is important: deterministic validation strengthens initial calibration without converting the semantic pipeline into a permanently intrusive query workload.

4.4 RAG artefacts and graph topology

Validated semantic schemas are transformed into atomic Markdown artefacts. The chunking strategy preserves complete semantic units rather than slicing text by arbitrary token counts. Separate artefacts describe schemas, tables, joins and operational context. Each chunk contains source identifiers that support evidence tracing from a generated answer back to the governed catalogue.

In parallel, the relational topology is exported to Amazon Neptune [14]. Nodes represent databases, schemas and tables, while edges encode validated relationships and associated properties such as key fields and confidence. This graph enables path finding across multiple hops and prevents the downstream agent from relying exclusively on vector similarity when selecting join routes.

4.5 Event-driven cloud operation

The production architecture separates local extraction from cloud processing. Change events are propagated through Kafka and Flink, whose distributed-log and stateful-streaming models support decoupled, replayable processing [9], [10]. Events are stored in Amazon S3 and processed asynchronously through EventBridge, queues and containerised workloads. This design absorbs bursts, isolates failures and supports idempotent retries. Infrastructure as code, least-privilege IAM policies and the AWS Well-Architected principles provide repeatable deployment and explicit security boundaries [15].

4.6 AXON continuous improvement loop

AXON begins with negative feedback associated with a real downstream interaction. A Locator agent identifies the affected semantic scope and an Architect agent diagnoses the likely root cause and proposes corrections. The system never applies these proposals directly. Changes are presented in a review portal, where a human can inspect file-level differences, approve or reject the proposal and trigger a controlled merge into production. The updated artefacts are then re-indexed, closing the loop between use and maintenance.

5. Evaluation methodology

5.1 Industrial environment

Evaluation was performed in a real steel-manufacturing environment containing three heterogeneous databases, 2,118 tables and 31,735 fields. The study therefore examines behaviour at catalogue scale rather than on a small curated demonstration. Evaluation combined structural pipeline metrics, downstream agent metrics, AXON operational outcomes, a regression-oriented golden dataset and qualitative case studies.

5.2 Aletheia evaluation

Aletheia was assessed through coverage, artefact generation, join filtering, graph construction and downstream usefulness. Coverage measured the proportion of physical catalogue entities represented in governed outputs. Artefact quality considered structural completeness, bounded chunk size, traceability and resistance to malformed model outputs. Relationship validation focused on the removal of incompatible or unsupported candidates before graph ingestion.

5.3 Downstream evaluation through Ariadne

Because Aletheia is infrastructure rather than an end-user application, part of its value was evaluated indirectly through Ariadne. The conversational agent retrieves Aletheia artefacts, reasons over the graph and produces grounded explanations and T-SQL proposals. Four dimensions were measured: answer relevance, correctness, faithfulness to retrieved context and transparency, following the multi-dimensional logic used by RAG evaluation frameworks rather than relying on a single aggregate score [12]. These metrics indicate whether the

semantic layer supports useful and traceable downstream behaviour, but they are not treated as an absolute ground truth for the catalogue itself.

5.4 AXON and golden dataset

AXON was evaluated on real user feedback and on a golden dataset derived from historical reviewed cases. The deterministic layer checks regression-oriented functional constraints, including whether context generation matches the required fix scope, whether the Architect respects immutable artefact boundaries by avoiding new schema or join chunks, and whether execution finishes with a well-formed output. A PASS indicates that all three contractual conditions are satisfied; a FAIL identifies a structural regression. An LLM-as-a-judge layer then assesses problem resolution, technical correctness and actionability. Consistent with current judge-based evaluation practice, this qualitative layer enriches rather than replaces the deterministic verdict [11].

The dataset is not an independent factory-native truth source. Its reference decisions originate from earlier human review and are therefore vulnerable to anchoring and catalogue bias. For this reason, the results are interpreted as regression evidence under an operational contract, not as a universal proof of semantic correctness.

6. Results

Indicator	Observed result
Databases / tables / fields	3 / 2,118 / 31,735
Markdown knowledge chunks	4,688
Indexable semantic content	12.48 million characters
Ariadne: relevance / correctness	97.29% / 88.72%
Ariadne: faithfulness / transparency	88.95% / 81.34%
AXON golden dataset	24/24 deterministic PASS
LLM judge: solved / correct / actionable	4.27 / 4.18 / 4.32 out of 5
AXON detected inconsistencies	39
AXON actionable proposals / production merges	36 / 35
Mean proposal-review similarity	0.9166
Audited AWS cost	USD 4,553.61
Aletheia / AXON pipeline cost	USD 2,517.52 / USD 10.36

Table 1. Main scale, quality, maintenance and cost indicators.

The system generated 4,688 Markdown chunks and more than 12.48 million characters of indexable structural knowledge. These outputs cover table, column, relationship and operational context information while preserving references to source artefacts. The graph topology complements these chunks with explicit connectivity across the catalogue.

Downstream results show high relevance (97.29%) and strong correctness (88.72%) and faithfulness (88.95%). Transparency is lower at 81.34%, indicating that explanations and references remain an area for improvement even when the generated answer is substantively useful. The gap illustrates why a single aggregate quality score would conceal important differences between answer utility and auditability.

AXON converted real feedback into maintainable semantic changes. Across the operational sample, 39 inconsistencies were identified and 36 actionable proposals were generated; 35 had been merged into production at the evaluation cut-off, while one remained pending deployment. The regression suite obtained 24 PASS results across 24 cases. Independently, the LLM judge scored problem resolution at 4.27/5, technical correctness at 4.18/5 and actionability at 4.32/5. These results provide complementary evidence: the deterministic layer detected contractual regressions, whereas the judge characterised semantic usefulness.

Human review also remained close to the generated proposals: the mean sequence similarity between AXON suggestions and reviewed versions was 0.9166, and 61 patches were accepted with similarity between 0.99 and 1.00. This indicates that most interventions refined rather than rewrote the proposed technical content. The FinOps analysis recorded a total audited AWS cost of USD 4,553.61, including USD 2,517.52 attributable to Aletheia and USD 10.36 to AXON; cost is therefore concentrated in the initial enrichment stage, while feedback-driven maintenance operates at low marginal cost.

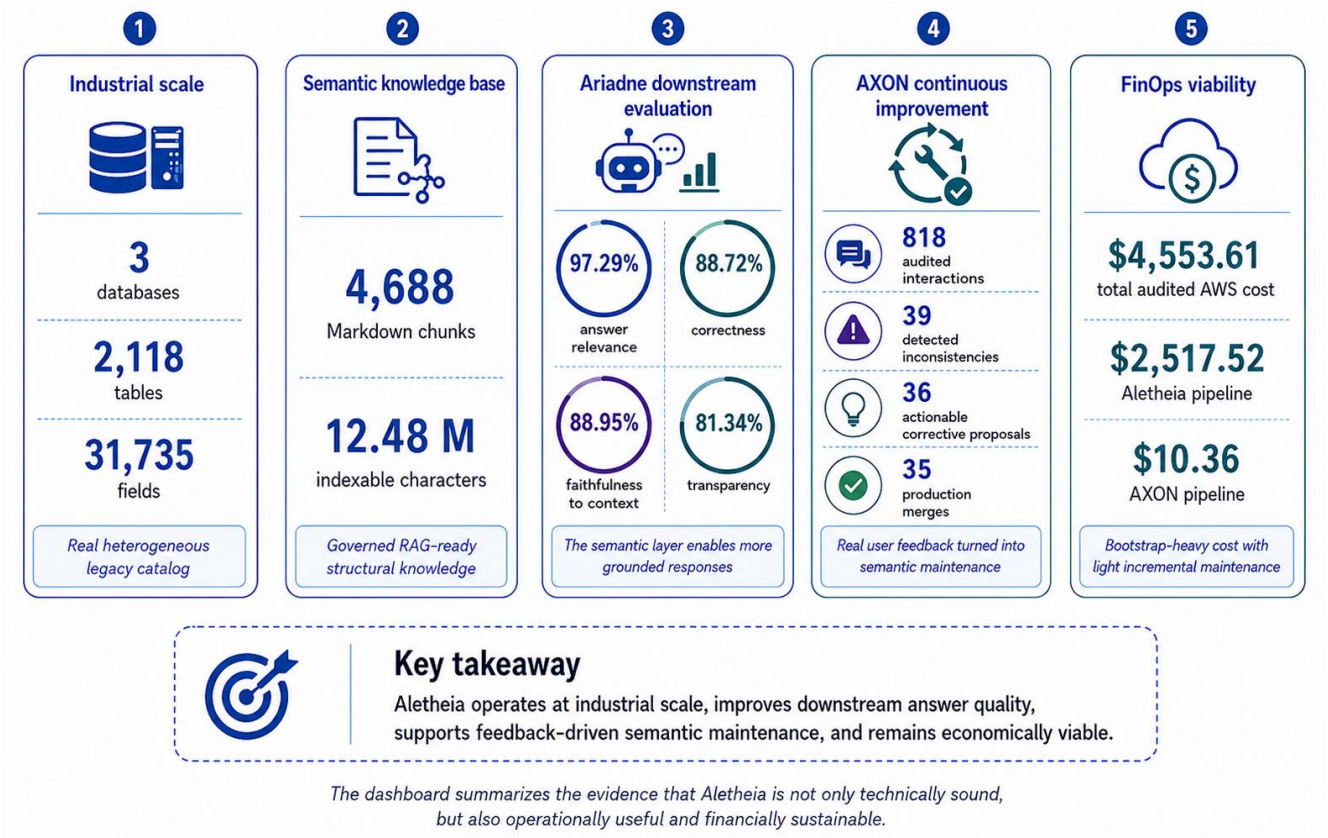

Fig. 2. Consolidated evaluation dashboard: industrial scale, downstream quality, continuous improvement and FinOps viability.

7. Discussion

7.1 Significance for intelligent industry

The results support the central hypothesis that reliable industrial generative AI depends on the quality of the knowledge layer rather than on model capability alone. Aletheia institutionalises meanings, constraints and relationships that were previously distributed across individual experts and legacy artefacts. The architecture therefore reduces organisational dependence on tacit knowledge and creates a reusable foundation for analysis, reporting and future decision-support agents.

The metadata-first choice is also strategically relevant. It allows an organisation to understand and govern its information landscape before opening direct autonomous access to production records. This sequencing reduces exposure, separates the operational and analytical planes and creates a clearer migration path towards governed data access or a future data-lake architecture.

7.2 Why hybrid validation is necessary

Neither probabilistic nor deterministic mechanisms are sufficient in isolation. LLMs are effective at inferring meaning from sparse and inconsistent naming, but they can produce unsupported yet plausible claims. Deterministic checks can verify types, existence and contracts, but they cannot fully judge business meaning. Aletheia therefore uses probabilistic enrichment for semantic coverage and structural controls for admissibility. AXON applies the same principle to maintenance by combining PASS/FAIL regression rules with a qualitative judge and human approval.

7.3 Limitations and threats to validity

The first limitation is the absence of a permanent multidisciplinary data-governance committee capable of exhaustively validating every generated description. Technical controls and developer review reduce risk, but they do not replace the process expertise of senior operators and metallurgical engineers. The semantic catalogue should therefore be regarded as governed and operationally useful, not infallible.

Second, the AXON golden dataset inherits decisions from historical human review. It is valuable for detecting regressions against the established operational contract, but it is not an independent absolute truth. Both deterministic rules and LLM judges may reproduce assumptions embedded in the original catalogue or review process.

Third, model stochasticity means that repeated executions may differ even under equivalent inputs. The current framework emphasises functional contracts and semantic plausibility, but larger repeated-run studies would be required to quantify variance and confidence intervals. Similarly, the downstream Ariadne metrics are based on a finite set of interactions and may evolve as prompts, models and user behaviour change.

Fourth, retrieval exhibits a structural precision–recall trade-off. Broad retrieval protects recall but can inject irrelevant context; aggressive filtering improves precision but may omit the relationship required for a multi-hop query. Future evaluation should report retrieval quality separately by artefact type and query complexity rather than relying on a single geometric measure.

Finally, the architecture depends on industrial connectivity, cloud inference quotas and controlled synchronisation between local and cloud components. Long communication outages delay catalogue freshness, while large bootstrap runs can encounter inference throttling. Radical simultaneous schema changes may also require manual conflict resolution in the review portal.

These constraints are acceptable under the current fail-safe design but limit claims of fully autonomous operation.

7.4 Representative qualitative cases

Quantitative indicators were complemented with representative cases because many catalogue defects are difficult to express through a single scalar metric. In the first case, an abbreviated mainframe table and its fields could not be interpreted from their physical names alone. The enrichment stage combined bounded value distributions, neighbouring columns and historical query evidence to infer the operational concept. The resulting description allowed the downstream agent to retrieve the correct source without requiring the user to know the legacy nomenclature. This case illustrates the value of contextual inference when traditional documentation is absent.

A second case concerned cross-system type incompatibility in a candidate relationship. Two fields represented the same business identifier, but one was stored as a character value and the other as a numeric value with formatting artefacts. The discovery stage identified high semantic similarity, while the validation and repair logic generated a cast-aware join instruction and documented the conversion. The relationship was therefore retained without hiding the physical mismatch from future consumers. This is preferable to either rejecting a valid business relation or silently assuming direct type compatibility.

A third case originated from negative user feedback that was initially classified as a catalogue problem. AXON reconstructed the conversation and determined that the requested concept was outside the available scope and that the generated answer had correctly communicated the limitation. The case was reclassified as USER_ERROR and no semantic files were modified. This behaviour is important because a feedback-driven system that interprets every dissatisfied interaction as a metadata defect will progressively corrupt a correct catalogue.

In a fourth case, a valid CHUNK_ERROR was traced to an incomplete treatment of null values in an operational field. The Locator selected the affected artefact, the Architect proposed a bounded correction and the review portal displayed the exact semantic difference. Once approved, the merge updated the catalogue and triggered re-indexing. The case demonstrates the intended granularity of AXON: local corrections are preferred to uncontrolled regeneration of entire schemas.

Finally, the enriched catalogue supported a fine-grained industrial analysis that required traversing multiple tables with different grains. The graph constrained the join path, while the retrieved Markdown artefacts exposed the relevant field meanings and aggregation warnings. The resulting query could be used to automate a recurring operational report. This case is relevant to MIINT because it links semantic infrastructure with concrete process intelligence rather than treating catalogue generation as an isolated documentation exercise.

7.5 Reliability and operational safeguards

The production design includes safeguards intended to transform model failures into observable technical events. All intermediate artefacts are persisted before subsequent processing, allowing failed stages to be replayed without restarting the complete pipeline. Queue-based decoupling absorbs bursts and protects source systems from cloud-side backpressure. Idempotency keys prevent duplicate events from generating inconsistent catalogue versions, and dead-letter queues isolate records that repeatedly violate contracts, consistent with reliable cloud architecture principles [15].

Model outputs are never treated as executable instructions. YAML responses must comply with explicit schemas, reference only known physical entities and pass post-inference filters. The graph loader consumes generated exports rather than direct free-form model text. Similarly, AXON proposals remain in a pre-production buffer until human approval. These barriers address recurring LLM application risks such as improper output handling, excessive agency and unvalidated generated content [16].

The architecture also distinguishes degradation from failure. If the on-premise communication channel becomes unavailable, change events can remain buffered and catalogue freshness is temporarily reduced, while the last validated semantic version remains available. If inference quotas are exceeded during a massive bootstrap, exponential backoff increases completion time but does not alter previously governed outputs. If a radical schema migration produces conflicting patches, the automatic merge blocks safely and requires interactive resolution. In each scenario, the system prioritises semantic consistency over immediate freshness.

7.6 Economic and organisational implications

The cost distribution suggests that semantic bootstrapping and semantic maintenance have different economic profiles. Initial enrichment requires processing the complete catalogue, generating embeddings and constructing all downstream artefacts. Incremental operation, by contrast, processes only changed entities or reviewed feedback. This difference explains why the Aletheia pipeline accounts for a substantial share of the audited expenditure while AXON remains inexpensive. For corporate adoption, the relevant comparison is therefore not only total cloud cost, but the recurring cost of maintaining equivalent documentation manually and the opportunity cost of expert dependency.

The architecture changes the role of domain experts rather than eliminating it. Without automation, experts must repeatedly explain the same schema details to analysts and developers. With Aletheia, their scarce time can be concentrated on validating ambiguous concepts, resolving high-impact conflicts and governing critical changes. The semantic layer captures validated knowledge so that it can be reused across users and applications. This creates an organisational memory that persists beyond individual availability and supports cross-plant transfer of analytical capabilities.

Nevertheless, successful deployment requires explicit ownership. A semantic catalogue cannot be maintained solely as an AI experiment; it needs defined responsibilities for source-system changes, review priorities, access policies and quality escalation. The technical loop implemented by AXON provides the mechanism, but organisational governance determines which proposals are reviewed, who is authorised to approve them and how conflicting interpretations are resolved.

7.7 Reproducibility and evaluation evolution

Reproducibility in a cloud-based generative system is necessarily bounded. Source artefacts, prompts, model identifiers, parameters and generated outputs are versioned so that an execution can be reconstructed. However, managed foundation models may evolve and stochastic sampling can produce different wording or rankings. For this reason, the evaluation framework focuses on stable functional contracts and preserved evidence rather than exact textual equality.

The next evaluation iteration should introduce repeated executions per golden case and report pass probability, judge-score dispersion and disagreement between deterministic and qualitative layers. Cases should also be stratified by error type, affected artefact, number of relational hops and required scope. This would distinguish easy local corrections from complex topology changes and prevent aggregate scores from being dominated by the most frequent category.

A further improvement would be the creation of an independent expert panel for a representative sample. Reviewers from data engineering, process operations and business analysis could score semantic correctness, usefulness and risk separately. Inter-rater agreement would reveal where the catalogue contains genuinely ambiguous concepts rather than simple model errors. Such a panel is expensive, but it would provide a stronger external reference than historical operational decisions alone.

8. Conclusions and future work

This paper presented Aletheia, a governed semantic layer that converts heterogeneous industrial legacy metadata into structured knowledge for generative AI agents. The system combines technical profiling, LLM enrichment, relationship discovery, deterministic quality gates, retrieval artefacts and graph topology. AXON extends the architecture with a human-supervised mechanism that turns real feedback into controlled corrections.

Evaluation at industrial scale demonstrates both technical feasibility and operational relevance. The generated knowledge base covers thousands of tables and tens of thousands of fields, improves downstream answer quality and can be maintained incrementally at low marginal cost. At the same time, the study shows that semantic quality cannot be reduced to a single automated metric: expert validation, regression contracts, qualitative assessment and transparent limitations remain essential.

Future work will extend the architecture towards governed access to production data, integration of real-time MQTT signals and deployment across additional plants. On top of this shared semantic layer, specialised agents could generate technical reports, construct process simulators, evaluate operational scenarios and provide justified recommendations under human oversight. The long-term objective is not autonomous industrial decision-making, but a corporate decision-support infrastructure grounded in traceable data and governed knowledge.

Declaration of competing interests

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The industrial data, catalogue artefacts and operational traces used in this study contain confidential corporate information and cannot be made publicly available. Aggregated results and architectural descriptions are reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the development of this work, generative AI tools were used as complementary support. They were mainly used to support early planning and synthesis, assist with software development, clarify technical questions, consult information related to AWS services, and improve the visual presentation of selected figures. Any suggestions produced by these tools were reviewed, adapted and checked against official documentation and the deployed system before being incorporated. The project requirements, architectural design, implementation, validation, interpretation of results and conclusions were developed and supervised by the author, who takes full responsibility for the content of this publication.

References

[1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[2] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[3] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv:2312.10997, 2023.

[4] D. Edge et al., “From Local to Global: A Graph RAG Approach to Query-Focused Summarization,” arXiv:2404.16130, 2024.

[5] A. Hogan et al., “Knowledge Graphs,” ACM Computing Surveys, vol. 54, no. 4, pp. 1-37, 2021.

[6] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of EMNLP-IJCNLP, pp. 3982-3992, 2019.

[7] S. Rose, O. Borchert, S. Mitchell and S. Connelly, Zero Trust Architecture, NIST Special Publication 800-207, 2020.

[8] International Society of Automation, ISA/IEC 62443 Series of Standards: Industrial Automation and Control Systems Security, 2026.

[9] J. Kreps, N. Narkhede and J. Rao, “Kafka: A Distributed Messaging System for Log Processing,” Proceedings of the NetDB Workshop, 2011.

[10] P. Carbone et al., “Apache Flink: Stream and Batch Processing in a Single Engine,” IEEE Data Engineering Bulletin, vol. 38, no. 4, pp. 28-38, 2015.

[11] L. Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” Advances in Neural Information Processing Systems, vol. 36, 2023.

[12] S. Es, J. James, L. Espinosa-Anke and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” Proceedings of EACL: System Demonstrations, pp. 150-158, 2024.

[13] Amazon Web Services, Amazon Bedrock Documentation, AWS Documentation, accessed July 2026.

[14] Amazon Web Services, Amazon Neptune Documentation, AWS Documentation, accessed July 2026.

[15] Amazon Web Services, AWS Well-Architected Framework, AWS Documentation, accessed July 2026.

[16] Open Worldwide Application Security Project, OWASP Top 10 for Large Language Model Applications, 2025.