Trustworthy Data Foundations for AI-Driven Analytics in Distributed IoT: A Validation-First Methodology

Maziar Ghorbani[1 [0000−0001−7284−9853]], Diana Suleimenova[1 [0000−0003−4474−0943]], Laura Harbach[1 [0000−0001−7944−0292]], Pramit Mazumdar[2], Niruban Paramanathan[3], Ricardo Severino[4 [0000−0002−4215−3238]], Özer Aydemir[5], and Derek Groen[1,6 [0000−0001−7463−3765]]

[1] Department of Computer Science, Wilfred Brown Building, Brunel University of London, Kingston Lane, Uxbridge, Middlesex, UB8 3PH, United Kingdom
maziar.ghorbani@brunel.ac.uk
[2] GetFudo Ltd, Covent Garden, London, United Kingdom
[3] Dealdio Ltd, Braintree Road, Ruislip, United Kingdom
[4] Institute for Systems and Computer Engineering of Porto, Portugal
[5] IOTIQ GmbH, Emilienstrasse 13, 04107 Leipzig, Germany
[6] Faculty of Science (Informatics Institute), University of Amsterdam, Science Park, 904 1098 XH, Amsterdam, Netherlands

Abstract. Reliable large language model (LLM)-assisted operations in distributed Internet of Things (IoT) systems depend fundamentally on the quality, structure, and provenance of telemetry available at inference time. In this paper, we present a validation-first methodology for AI-driven analytics that enforces end-to-end verification of the data pipeline before permitting LLM inference and retrieval-augmented generation. The approach formalises three pre-inference quality gates addressing infrastructure and schema conformance, data integrity across freshness and continuity constraints, and context readiness through deterministic, ID-bound prompt construction. This methodology is implemented in HOMEPOT (Homogeneous Cyber Management of End-Points and Operational Technology), a unified endpoint and IoT management platform supporting heterogeneous devices and MQTT-connected sensors. We conducted a 10-day Technology Readiness Level (TRL-4) pilot involving 10 devices, processing over 140,000 telemetry samples and health checks alongside continuous state-transition logging. Integrity indicators demonstrate 100% completeness for key telemetry fields and sub-minute maximum inter-arrival gaps, confirming strong temporal continuity. These results establish measurable readiness conditions under which AI inference is treated as a conditional capability rather than an assumed default, providing a reproducible blueprint for dependable AI integration in distributed IoT environments rapidly transitioning toward pilot-scale deployment.

Keywords: Validation-first methodology · Distributed IoT systems · Data quality assurance · Retrieval-augmented generation · Trustworthy AI · Telemetry integrity

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

1 Introduction

Generative AI (GenAI) and large language models (LLMs) are increasingly being adopted as operational copilots for monitoring, diagnosing, and optimising distributed Internet of Things (IoT) and edge-device fleets. In practice, these assistants are frequently asked to produce *actionable* guidance under uncertainty (e.g., "why are devices flapping?", "which sites are at risk?", "what should we fix first?"). However, the dependability of LLM-mediated operations is limited by the quality, freshness, and structure of the underlying data context. In this work, we functionally define "trustworthiness" not as absolute cognitive truth from the LLM, but as *deterministic data readiness*, providing strict bounds on input uncertainty before any reasoning occurs. When telemetry is incomplete, delayed, inconsistent across sources, or assembled into prompts without explicit provenance, LLMs tend to produce confident but incorrect narratives. This represents a significant failure mode in operational settings where the cost of wrong actions is high [1,2].

Despite extensive research on anomaly detection, observability, and AI for operations [3,4], many industrial IoT deployments still exhibit production anti-patterns in which data pipelines are deployed before systematic validation, leading to downstream discovery of missing fields, schema mismatches, timestamp drift, or noisy duplication [5]. These issues are not merely engineering inconveniences; they directly affect downstream analytics and AI layers by corrupting the evidence presented inside the model's context window. As a result, standard LLM grounding strategies, such as retrieval-augmented generation (RAG), can inadvertently amplify these flaws. For example, retrievals can return stale or irrelevant "memories", whilst missing or malformed real-time signals force the model to fill gaps with plausible-sounding guesses [6,1]. This poses a potentially significant issue as, with the growing integration of GenAI, such platforms could eventually generate code, security policies, predictive models, and complex behaviours from large streams of validated data.

This paper argues that reliable AI-driven analytics in distributed IoT systems requires a *validation-first* approach: the pipeline must be verified before the inference is trusted. Here, we propose and implement a methodology that treats data quality and context readiness as first-class, testable artefacts. We apply this approach within the HOMEPOT project, a unified IoT/endpoint management platform currently validated at TRL-4, with the present pilot study designed to evaluate readiness for progression towards TRL-5.

HOMEPOT (Eureka - ITEA 4 Project) was developed in collaboration with industry partners as part of a staged, requirements-led programme. Initial work focused on consolidating operational requirements, specifications, and state-of-the-art constraints, followed by a second phase that focused on data analytics and AI integration. As a result, the system features studied in this paper are not ad hoc prototype additions; they are deliberately designed to satisfy externally driven requirements for operational visibility, traceability, and explainable decision support.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

The primary contributions of this work are threefold. First, we introduce a validation-first protocol that formalises pre-inference enforcement through three sequential quality gates. These gates operate as a pipeline: establishing the connection infrastructure and metric contracts (Gate A), validating deterministic data continuity and integrity thresholds (Gate B), and finally verifying structural context readiness and ID-bounding before the LLM consumes the prompt (Gate C). Second, we present the HOMEPOT platform as an architectural instantiation of this protocol, integrating time-series telemetry, persistent alert identifiers, and a grounded cognitive engine. Third, we provide a reproducible TRL-4 case study demonstrating quantifiable integrity indicators that define measurable readiness conditions under which AI inference is permitted.

The remainder of this paper is structured as follows. Section 2 reviews background and related work on IoT management, data-centric quality practices, and trustworthy GenAI grounding. Section 3 describes HOMEPOT's data analytics architecture and the data foundations designed for AI consumption. Section 4 presents the validation-first protocol and its quality gates. Section 5 details the grounded cognitive engine that converts validated data into diagnostic explanations. Section 6 reports the case-study setup and evaluation results. Section 7 concludes with limitations and future work.

2 Background and Related Work

To establish the theoretical basis for a validation-first methodology, HOMEPOT is positioned at the intersection of AIOps, data-centric quality frameworks, and grounding mechanisms for generative AI. This approach ensures that AI-driven diagnostics operate only on verified evidence to mitigate operational risks.

Traditional operational monitoring relies on statistical and machine learning techniques to detect abnormalities [3]. Although log- and metric-driven techniques can identify abnormal behaviour and support diagnosis [7], their effectiveness depends fundamentally on upstream data integrity. Issues such as missing fields, duplicated events, inconsistent identifiers, or timestamp drift distort signals and degrade downstream analytics [4].

Recent shifts toward data-centric AI argue that trustworthiness extends beyond accuracy to include completeness, consistency, and timeliness [2]. Improvements in data collection and validation often outweigh incremental model refinements [8]. In IoT contexts, this motivates explicit validation of telemetry pipelines, including freshness, continuity, and range enforcement before higher-level analytics or AI reasoning are applied [5].

The integration of LLMs introduces additional risks, as surveys of hallucination and factuality show that models can generate fluent but incorrect outputs when evidence is incomplete or weakly structured [6,9], and scaling alone does not guarantee truthful behaviour [10]. In operational settings, such confident misstatements can lead to inappropriate or unsafe interventions, making an evidence-grounded explanation essential.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

Retrieval-augmented generation seeks to improve reliability by conditioning outputs on retrieved evidence [11], often implemented via dense retrieval over historical incidents or vector stores [12]. However, retrieval cannot compensate for stale or malformed real-time context, since retrieval-based grounding remains bounded by the quality and structure of the underlying corpus [6,10].

Although substantial advances exist across anomaly detection, data quality, and retrieval-based grounding, many systems lack an explicit interface between telemetry validation and LLM context construction. This enables a deploy-first pattern in which AI operates over partially validated evidence [4]. Such patterns conflict with emerging AI risk management guidance that emphasises governance, traceability, and verifiable evidence in AI-enabled systems [13].

Taken together, the absence of an explicit interface between upstream telemetry validation and downstream LLM context construction enables deploy-first configurations in which inference may proceed over partially validated evidence. This architectural gap motivates the validation-first framework introduced in the next section.

3 System Architecture and Data Foundations

During ordinary operation, a diverse set of endpoints (such as managed hardware, web clients, and IoT sensors) and software bridge agents stream telemetry and lifecycle events to a centralised FastAPI backend. As illustrated in Fig. 1, the HOMEPOT architecture follows a layered pipeline from distributed devices and dashboards through API and service layers to the data layer (PostgreSQL, TimescaleDB, ChromaDB) and the grounded cognitive engine. To preserve architectural integrity and data sovereignty within this layered design, the AI subsystem operates as a native, local-first module, forming a clearly bounded inference layer built upon auditable data foundations.

3.1 Operational Data Model as LLM-Ready Context Sources

To facilitate both high-frequency monitoring and forensic auditing, the platform partitions operational evidence into distinct schema-enforced tables. The health and time-series telemetry layer utilises TimescaleDB hypertables to accelerate range queries, capturing devices' health status and response latency alongside extensible payloads. Parallel to these metrics, the system documents lifecycle transitions and job outcomes within dedicated history tables to provide a granular trace of device behaviour. Governance and provenance are maintained through a multi-layered logging strategy where middleware-driven request logs intersect with categorised error metadata and explicit audit trails. Finally, the platform generates persistent alert records with unique identifiers that serve as stable, immutable references for subsequent AI-driven explanations and operator intervention.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

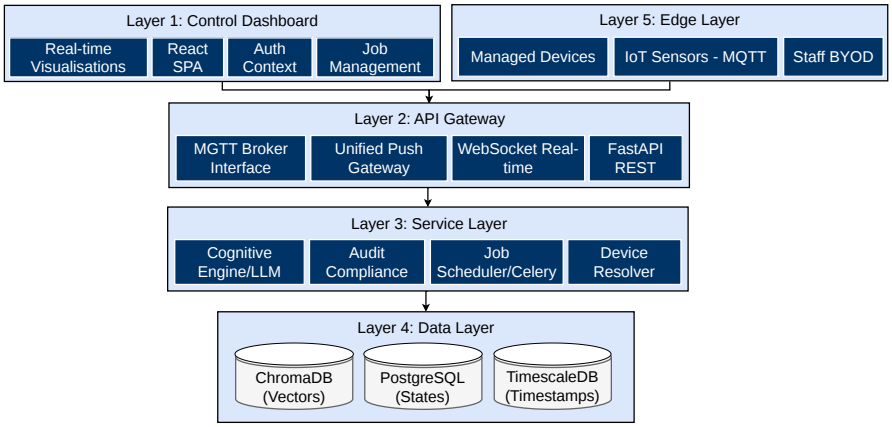

Fig. 1. Layered architecture of the HOMEPOT platform, illustrating the end-to-end pipeline from heterogeneous IoT devices and dashboards through the API and service layers to the data layer (PostgreSQL, TimescaleDB, ChromaDB) and the cognitive engine. The design emphasises auditable data flow and local-first AI integration.

3.2 Time-Series Optimisation and Queryability

Ensuring the low-latency retrieval of operational evidence necessitates the deployment of specialised TimescaleDB management utilities. The system automatically partitions health-check data into weekly chunks and implements aggressive compression and retention policies to maintain performance over long durations. For complex analytics, the system uses continuous aggregates to pre-compute hourly and daily summaries by decoupling real-time ingestion metrics from the intensive computational requirements of site-level rollups and device performance summaries.

3.3 Noise Control: Smart Data Filtering

Operational telemetry can overwhelm storage and hinder trustworthiness if it contains redundant snapshots or noisy jitter that dominates analytics. The system, therefore, implements *smart filtering* for device metrics. The architecture guarantees baseline snapshot is stored every 5 minutes for continuity, but triggers immediate storage only upon significant metric fluctuations or system restarts. This approach preserves high-resolution traces during anomalous events while maintaining a predictable baseline for routine monitoring.

3.4 LLM Context Interfaces

A core tenet of the architecture is the treatment of database tables as deterministic context sources governed by a stable formatting contract. Rather than passing free-form logs to the LLM, the AI query endpoint assembles a

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

structured context document that prioritises real-time system state and active alerts over historical artefacts. This deterministic assembly enforces explicit identifiers, bounded sections, and consistent ordering, ensuring that the LLM grounds its explanations in referenceable evidence rather than unstructured telemetry. By constraining context construction at the source, the system reduces the risk of fabricated identifiers or unsupported system states.

4 Validation-First Methodology (Protocol)

The validation-first methodology is built on a single architectural principle: AI inference is treated as a *downstream consumer* of operational data and must therefore be gated on the readiness of the upstream pipeline. This principle is operationalised as a tiered verification envelope consisting of three sequential quality gates, as illustrated in Fig. 2, which can be executed repeatedly during development and before production or pilot rollouts. Failure to satisfy any individual gate triggers a constrained non-actionable mode that precludes unrestricted AI execution, preventing the generation of narratives based on unstable or incomplete evidence.

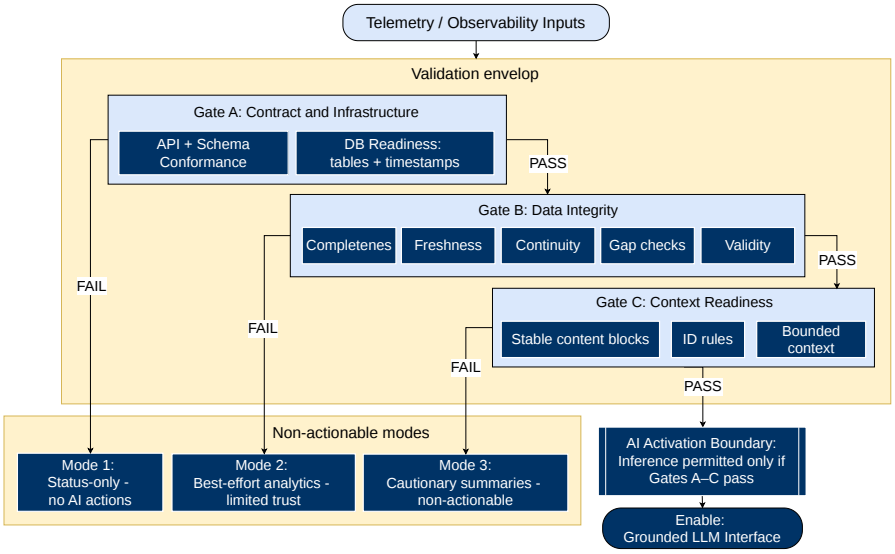

Fig. 2. Validation-first protocol structured as a control sequence. Telemetry must satisfy three sequential gates: Contract and Infrastructure (Gate A), Data Integrity (Gate B), and Context Readiness (Gate C), before crossing the AI activation boundary that permits grounded LLM inference. Failure at any gate triggers a defined non-actionable mode with constrained behaviour rather than unrestricted AI execution.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

4.1 Gate A: Infrastructure and Contract Verification

Gate A validates that the data-producing interfaces are stable and schema-conformant. The system uses Pydantic schemas within FastAPI to enforce strict range constraints and required fields at the point of ingestion. This strategy shifts data quality responsibilities to the interface level, ensuring that subsequent analytics operate on a verified foundation. Furthermore, the protocol includes automated schema inspection, endpoint contract validation, audit-trail verification, and database readiness checks, specifically confirming that TimescaleDB hypertables are prepared to sustain anticipated time-series workloads. If Gate A fails (e.g., schema mismatch, missing tables, broken API paths), AI inference is treated as unsafe and is blocked or degraded because any subsequent analytics would be built on unstable evidence.

4.2 Gate B: Data Integrity Assurance

Gate B evaluates whether operational evidence satisfies the minimum integrity thresholds required for reliable analytics and grounded LLM inference. While Gate A guarantees structural and contractual correctness at ingestion, Gate B verifies that the accumulated telemetry is sufficiently complete, continuous, and plausible to support dependable reasoning.

An automated validation suite monitors collection rates over defined observation windows and detects temporal discontinuities that could distort anomaly detection or trend analysis. Freshness checks ensure that core analytics tables are actively receiving data, while continuity checks identify gaps that exceed predefined tolerances. These safeguards prevent the system from deriving conclusions from stale or sparsely sampled telemetry.

In addition, Gate B enforces plausibility and completeness constraints at the metric level. The system flags null values for critical fields and detects physically implausible observations, such as resource utilisation percentages exceeding logical bounds. This ensures that downstream analytics and explanatory narratives are grounded in evidence that is both internally consistent and operationally credible.

When freshness or continuity requirements are not met, for example, due to extended collection gaps or empty critical tables, the cognitive engine reverts to best-effort status reporting and explicitly marks generated narratives as not audit-ready. In this mode, the system may summarise an observable state but does not present high-confidence explanations or recommendations, thereby preventing overinterpretation of incomplete evidence.

4.3 Gate C: Context Readiness for LLM Ingestion

Even with correct and timely data, LLM grounding can fail if operational evidence is assembled without a stable structure or explicit identifier conventions. Gate C evaluates whether the system can generate a deterministic, bounded, and referenceable context document suitable for grounded inference.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

This gate enforces stable section headers, strict identifier integrity, and consistent ordering to ensure that alerts and system states are referenced using explicit IDs. Role constraints and token discipline further limit fabrication by clearly separating the current operational state from long-term semantic memory.

If Gate C fails to produce a structurally valid context document, for example, due to missing required blocks, absent alert identifiers, or uncontrolled context growth, the system suppresses action-oriented recommendations and instead returns limited summaries that explicitly communicate uncertainty.

Table 1. Summary of Validation-First Quality Gates and Passing Conditions.

Quality Gate	**Objective**	**Passing Condition / Threshold**
Gate A: Infrastructure	Ensure connection and schema match contracts.	API endpoints responsive; device metrics conform strictly to expected JSON/SQL schema.
Gate B: Data Integrity	Verify data completeness, continuity, and freshness.	100% non-null key fields; timestamp drift $\leq \pm 5$s; inter-arrival gaps ≤ 60s (threshold).
Gate C: Context	Assure structural readiness for the LLM prompt.	Complete assembly of context blocks; strict ID referencing; tokens bounded within limits.

5 Grounded Cognitive Engine

The cognitive engine executes a grounded reasoning workflow, positioning LLM generation as an interpretive layer superposed upon validated operational evidence. As illustrated in Fig. 3, the architecture synthesises deterministic detection mechanisms with structured context assembly and retrieval-augmented memory. To facilitate natural language interrogation, the system aggregates context from its assembly blocks.

These blocks comprise real-time system status retrieved from PostgreSQL, including site and device metrics, push-notification statistics, and active alerts; long-term semantic memories retrieved from the vector store as prior incident resolutions; bounded short-term conversation history; and static knowledge derived from a real-time recursive scan of codebase documentation and structure. Together, they form a contiguous prompt context partitioned by explicit block headers, enabling verifiable alignment between LLM outputs and underlying evidence.

The platform adopts a hierarchical memory architecture in which persistent ChromaDB collections store long-term vector embeddings. Conversely, short-term conversational state is truncated to the most recent exchange cycles, ensuring the prompt remains strictly within the bounds of the model's maximum token capacity.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

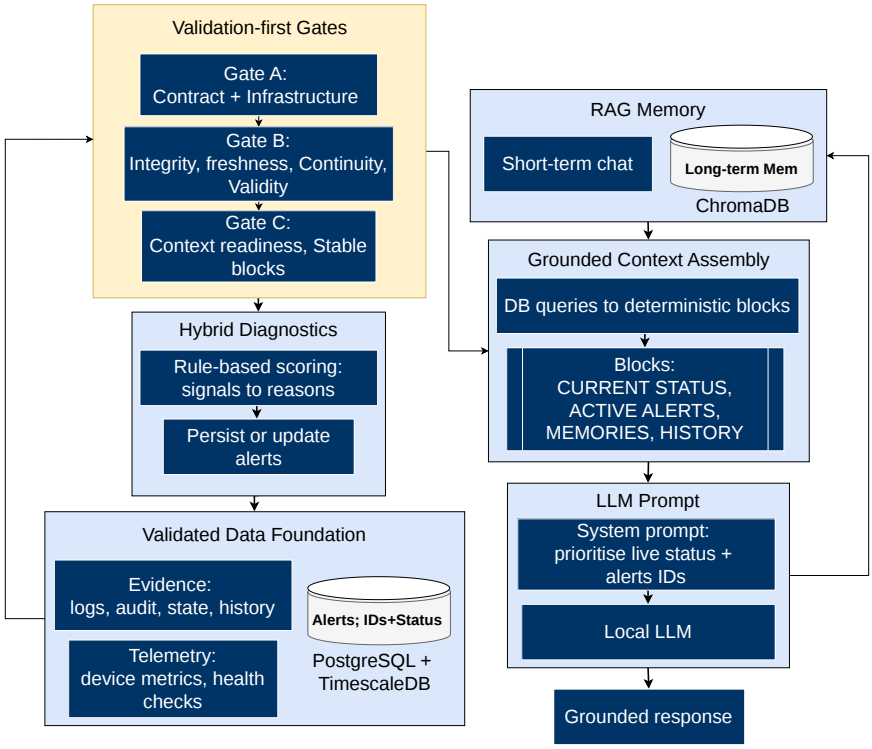

Fig. 3. Cognitive engine integrating validation-first gates, hybrid anomaly detection, structured context assembly, and retrieval-augmented memory. Validated telemetry and persistent alert IDs are assembled into deterministic context blocks before local LLM inference, enabling evidence-checkable explanations or recommendations.

5.1 Query and Anomaly Workflows

Parallel to the natural-language query interface, the system operates a deterministic anomaly detection tier that computes quantitative scores and diagnostic signals from structured telemetry, including resources (e.g., CPU and Memory) utilisation and consecutive health-check failures. When an anomaly is detected (e.g., Devices "flapping", rapidly disconnecting and reconnecting or sudden drops in polling frequency), the architecture prioritises the creation or update of persistent alert records, ensuring that incidents are represented through immutable identifiers and explicit lifecycle states. The LLM does not independently infer anomaly status; rather, it conditions its diagnostic explanations and remedial recommendations on both the computed anomaly metrics and the current system context. This separation of quantitative detection from generative interpretation preserves evidential traceability while enabling natural-language reasoning over validated alerts.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

5.2 The Grounding Contract

The primary architectural safeguard of the cognitive engine is a grounding contract embedded in both context construction and the governing system prompt. Rather than relying on implicit alignment, this contract formalises enforceable constraints that shape how the LLM interprets operational evidence.

First, it establishes current-state primacy: the `current system status` block is treated as the authoritative source for all operational assertions. Second, it enforces identifier discipline by requiring explicit alert IDs in incident references and prohibiting the fabrication or inference of non-existent system identifiers. Third, it defines an uncertainty protocol that obliges the model to acknowledge evidential gaps instead of extrapolating unsupported system states.

In doing so, Gate A contract operationalises validation-first principles at the generative layer. Gate B ensures that current-state telemetry is fresh and continuous, while Gate C guarantees that contextual inputs are structurally bounded and referenceable. To illustrate Gate C, a simplified snippet of an assembled context enforcing grounding is shown below:

```
[SYSTEM PROMPT]
 Act as an IoT diagnostic assistant. Ground claims in EXACT IDs
provided under current status. Do not hallucinate system metrics.
[CURRENT STATUS]
 Device ID: DEV-991 | Status: WARNING | CPU: 95% | MEM: 75%
 Active Alert ID: ALT-042 | Category: Performance Exceedance
[END STATUS]
```

This explicit formatting ensures the LLM is constrained to reason within verified evidential limits, transforming the generation process from open-ended synthesis into a controlled interpretation of validated operational data.

6 Case Study and Evaluation

We evaluate HOMEPOT in a simulation-driven TRL-4 setting intended to support a transition toward TRL-5 initial pilot deployments. This assessment aims to demonstrate that validation-first gates yield quantifiable data-quality indicators and that the cognitive engine functions using referenceable, structured evidence. The trace comprises a heterogeneous set of 10 simulation devices (Linux/Windows/iOS endpoints and MQTT-based sensors) reporting health checks, performance metrics, state transitions, and operational logs to a FastAPI backend. For generative tasks, the cognitive engine was instantiated using the `Ollama 3.2` model (Ollama, Inc.) to provide a standard performance baseline.

To systematically test the gates, network anomalies such as missing packets and timestamp jitter were artificially injected during telemetry generation. Furthermore, the smart-filtering pipeline was configured to enforce a baseline collection rate of 288 daily snapshots per device (a 5-minute heartbeat); however, threshold-busting anomalies immediately override the filter to trigger accelerated

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

capture. Unless otherwise stated, reported counts in this paper are computed over a fixed query window (last 10 days), while continuity is computed over the last 7 days for health checks to reflect recent operational behaviour. The selection of this case study is due to an industry-led co-design process, which ensures that telemetry and logging features address their specific requirements. Consequently, the observed pipeline behaviours reflect stakeholder-driven operational demands rather than synthetic benchmarks.

6.1 Case Study Methods and Measurement Protocol

Measurement reporting is categorised into three classes. Volume and coverage assessments capture record counts across core tables and verify distinct device participation. Integrity indicators evaluate non-null completeness, detect validity violations, and quantify continuity gaps by measuring maximum inter-arrival times. Finally, collection-rate alignment compares observed sampling frequencies against the smart-filtering heartbeat baseline of daily snapshots. To ensure auditability and reproducibility, the system computes each quantitative metric via a standardised suite of SQL queries (Q1–Q8), summarised in Table 2, executed against the PostgreSQL instance. The resulting trace volume and coverage statistics are summarised in Table 3, while integrity and continuity indicators are reported in Table 4.

Table 2. Summary of reproducibility queries (Q1–Q8) used to compute quantitative indicators. Full SQL scripts are available in the project repository.

Query ID	Purpose
Q1	Volume and coverage counts (core tables)
Q2	Distinct device coverage
Q3	Field-level non-null completeness
Q4	Validity violations (range checks)
Q5	Global health-check continuity
Q6	Per-device health-check continuity
Q7	Device-metric continuity and gap detection
Q8	Collection-rate alignment under smart filtering

6.2 Measuring Dependability and Data Readiness

While absolute truthfulness from generative AI is not directly measurable, functional dependability can be inferred from observable properties of system behaviour. In this evaluation, quantitative indicators provide operational evidence for data readiness by capturing coverage, integrity, continuity, and collection alignment. Volume and coverage metrics demonstrate that analytics operate over the full device fleet rather than a selective subset. Completeness and validity checks provide evidence of internal data integrity, while continuity

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

measures reflect freshness and temporal stability. Collection-rate alignment further validates that the smart-filtering mechanism maintains a predictable baseline while capturing higher-fidelity traces during significant events. Together, these indicators define measurable preconditions under which inference can be considered operationally safe.

6.3 Case Study Results

We compute record counts for core tables over the evaluation window (Q1) and the number of distinct devices contributing to device metrics and health checks (Q2). These results substantiate the claim that the platform has accumulated operational evidence at a sufficient scale for preliminary analytics. Regarding completeness and validity, we also compute field-level non-null completeness for CPU/memory/disk/latency in `device_metrics` (Q3) and count validity violations for per cent-like metrics above 100% (Q4). As summarised in Table 3, the evaluation window captures full fleet participation and substantial telemetry volume across core operational tables. The corresponding integrity and continuity indicators are reported in Table 4.

In the current trace snapshot, completeness is 100% non-null for device metrics over the window (142,403 rows), with no observed CPU, memory, or disk utilisation values exceeding 100%. To assess continuity, we compute the maximum observed inter-arrival gap for `health_checks` over the last 7 days (Q5) and the per-device maximum gap distribution (Q6), ensuring that continuity is not dominated by a subset of devices. The global maximum inter-arrival gap is 0.526 minutes, while the maximum per-device gap reaches 0.593 minutes across the 10 devices. We additionally compute the number of gaps exceeding 60 minutes (a sustained discontinuity threshold) and the maximum gap over the evaluation window (Q7). In the current snapshot, we observe no inter-arrival gaps exceeding 60 minutes and a maximum inter-arrival gap of 0.524 minutes for device metrics. Finally, we consider expected collection rates under smart filtering. We compute observed snapshots per day for `device_metrics` by device and day (Q8) and compare these to the configured heartbeat baseline of 288/day, corresponding to a minimum expected rate under a 5-minute interval. Observed counts are substantially higher ($\approx$14.2k samples/device/day on the day with data), indicating that additional sampling triggers, such as significant-change updates or higher-frequency emitters in simulation, dominate the baseline. This illustrates the role of the baseline as a lower bound for continuity rather than a fixed-rate assumption.

When heterogeneous endpoints submit their diagnostic payloads, the backend first sanitises and structures the data into the operational database. The AI layer is then only invoked over this stabilised, queryable foundation. This isolation strategy not only protects sensitive operational metadata and preserves data sovereignty but also structurally eliminates the risk of the model hallucinating system states based on malformed or incomplete inputs, anchoring all downstream inferences firmly within verifiable reality.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

Table 3. Trace volume and coverage summary (computed from Q1–Q2 over the last 10 days).

Signal	Count
Distinct devices (metrics, health)	10
Device-metric samples (`device_metrics`)	142,403
Health checks (`health_checks`)	142,403
Error logs (`error_logs`)	69,999
API request logs (`api_request_logs`)	19,702
State transitions (`device_state_history`)	39,408

Table 4. Integrity indicators from Q3–Q7. Percentages are over the last 10 days; health-check continuity is over the last 7 days.

Indicator	Value
Non-null completeness (CPU, memory, disk, latency)	100%
Validity violations ($> 100\%$) (CPU, memory, disk)	0
Max inter-arrival gap (health checks, global)	0.526 min
Max inter-arrival gap (health checks, worst device)	0.593 min
Gaps > 60 min (device metrics)	0
Max inter-arrival gap (device metrics, global)	0.524 min

6.4 Case Study Findings and Qualitative Ablation

The validation-first workflow operates as a critical engineering control rather than a passive monitoring layer. Gate A surfaces interface and schema regressions before analytics and AI components are trusted; Gate B provides actionable diagnostics when integrity indicators degrade, including continuity gaps and staleness; and Gate C ensures that the LLM receives a bounded, referenceable context anchored to explicit incident identifiers. Compared to a deploy-first configuration, where analytics and generative explanations operate directly over whatever telemetry is available, the validation-first regime enforces evidence readiness prior to scoring and reasoning.

During early validation runs, we performed a qualitative ablation to compare the system's output with and without Gate B/C active. When the gates were bypassed, and incomplete telemetry representing a disconnected sensor was fed to the LLM, the model hallucinated plausible but unsupported root causes (e.g., claiming "high CPU usage caused thermal shutdown" due to stale token associations from vector-store memory). When the validation routing was engaged, Gate B immediately flagged the payload's timestamp staleness. The resulting Gate C prompt subsequently contained a strict empty-state alert, constraining the LLM to output: "Insufficient fresh telemetry available. Device disconnected at [Timestamp]; no valid CPU metrics for diagnostic reasoning." This ablation visibly illustrates how the framework prevents "garbage-in/garbage-out" failure modes by replacing ungrounded extrapolations with deterministic uncertainty bounds.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

More broadly, this shifts AI from an always-on diagnostic layer to a conditionally enabled component whose authority is contingent on measurable data integrity indicators. In doing so, the system reframes validation as a prerequisite operational state that emerges only when defined readiness conditions are satisfied.

7 Conclusions, Limitations and Scalability

This research establishes that dependable AI integration within distributed IoT is primarily a data-foundations challenge. Without rigorous telemetry validation, stable identifier management, and deterministic context assembly, LLM-based diagnostics inevitably produce ungrounded narratives. We presented a validation-first methodology implemented in HOMEPOT that operationalises this stance via three repeatable quality gates. Gate A enforces contract and infrastructure integrity, Gate B guarantees data freshness and continuity, and Gate C ensures context readiness for structured ingestion. This framework enables a cognitive engine that synchronises deterministic anomaly signals with persistent alert identifiers via a structured prompt contract, enabling explanations that can be explicitly checked against the underlying trace.

While this evaluation demonstrates feasibility and quantifiable integrity indicators, absolute claims of complete "trustworthiness" must be contextualised within this TRL-4 simulation phase. Several limitations remain regarding scalability and generalisation. First, the trace originates from a contained 10-device simulation and may not fully capture the behavioural heterogeneity and noise levels of production-scale IoT deployments. Second, the evaluation window is temporally constrained and therefore does not yet reflect long-duration operational drift. Third, the passing thresholds for freshness and continuity (e.g., 60-second continuity gaps) reported in this pilot are specifically calibrated to HOMEPOT's current baseline; scaling to distinct deployments will require dynamic, per-domain threshold tuning.

Future work will prioritise TRL-5 pilots with industry partners to evaluate the protocol under real operational stress and scale up heterogeneous device failures. We aim to incorporate a cybersecurity-oriented validation layer to strengthen data provenance through identity assertions. Lastly, the automation of table generation from direct SQL outputs will ensure reproducible consistency between the underlying database traces and reported results.

Acknowledgments. This work was carried out within the HOMEPOT (Homogeneous Cyber Management of End-Points and Operational Technology) project, an ITEA4-funded initiative supported by national funding agencies. The authors gratefully acknowledge the contributions of all consortium partners for their technical input and collaborative support.

Disclosure of Interests. The authors declare that they have no competing interests relevant to the content of this article.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42

Script Availability. To ensure reproducibility, the SQL script (`ICCS-501-paper-metrics.sql`) corresponding to queries Q1–Q8 and its accompanying reproduction guide (`ICCS-501-paper-metrics-reproduction.md`) have been formally committed to the HOMEPOT source repository (scripts directory). As HOMEPOT is an active industrial consortium project containing sensitive architectural infrastructure, full public release of the repository remains subject to cybersecurity clearance from our partners. However, the evaluation scripts and validation logic specifically accompanying this paper have been isolated to facilitate reproducibility: https://github.com/brunel-opensim/homepot-client.

References

1. Ni, B., Liu, Z., Wang, L., et al.: Towards trustworthy retrieval augmented generation for large language models: A survey (2025)
2. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems (1996). https://doi.org/10.1080/07421222.1996.11518099
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection. ACM Computing Surveys **41**(3) (2009). https://doi.org/10.1145/1541880.1541882
4. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: Hidden technical debt in Machine Learning systems. In: Advances in Neural Information Processing Systems (2015)
5. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the Big Data era. Data Science Journal (2015)
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys (2023). https://doi.org/10.1145/3571730
7. Du, M., Li, F., Zheng, G., Srikumar, V.: Deeplog. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017). https://doi.org/10.1145/3133956.3134015
8. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems (2009). https://doi.org/10.1109/MIS.2009.36
9. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.173
10. Lin, S., Hilton, J., Evans, O.: Truthfulqa: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (2022). https://doi.org/10.48550/arXiv.2109.07958
11. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Advances in Neural Information Processing Systems (2020). https://doi.org/10.48550/arXiv.2005.11401
12. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for Open-Domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020). https://doi.org/10.48550/arXiv.2004.04906
13. National Institute of Standards and Technology: Artificial intelligence risk management framework (AI RMF 1.0). Tech. Rep. NIST AI 100-1, National Institute of Standards and Technology (2023). https://doi.org/10.6028/NIST.AI.100-1

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_42