International Journal Software Engineering and Computer Science (IJSECS)
6 (3), 2026, 1058-1065
Published Online December 2026 in IJSECS (http://www.journal.lembagakita.com/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v6i3.7875.

RESEARCH ARTICLE Open Access

Integration of Static Knowledge and Telemetry Data for an IoT-Based Customer Support AI Agent

Jauhar Maknun Adib [1*], **Wahyu Purnama Magribi** [2], **Muhammad Fazly Qusyairy** [3], **Eric Julianto** [4], **Khusnul Fajri Rhomadon** [5]

[1*,2,3,4,5] Department of Computer Science, Esa Unggul University, West Jakarta City, Special Capital Region of Jakarta, Indonesia.

Article History:

Received: July 5, 2026.
Revised: July 8, 2026.
Accepted: August 27, 2026.
Available online: September 20, 2026.
Published: December 1, 2026.

***Corresponding Author:**
adibmaknun16@gmail.com

Abstract: The operational effectiveness of artificial intelligence agents in customer support is frequently attributed to the generative sophistication of underlying language models, although answer reliability primarily depends on the structure and quality of the referenced knowledge base. In complex technical domains such as the Internet of Things (IoT), static documentation alone often fails to resolve customer inquiries that involve real-time device states, sensor readings, and connectivity logs. This study investigates how progressive tiers of knowledge integration affect the response quality of an AI agent within an IoT customer support context. Employing a mixed-method pilot design combining literature synthesis, comparative analysis, experimental testing, and conceptual exploration, thirty representative questions were classified into informational, status, and diagnostic categories. These queries were evaluated across four operational scenarios: without retrieval-augmented generation (RAG), article-based RAG, article plus device metadata records, and a comprehensive hybrid framework integrating static articles, device records, and telemetry streams. Answer quality was measured using Cosine Similarity, BERTScore (F1), and a randomized blind human evaluation assessing relevance, correctness, and usefulness. The results demonstrate that while curated static documentation adequately addresses procedural informational queries, resolving status and diagnostic issues necessitates real-time operational context. The hybrid integration delivered the highest overall performance, achieving a human-evaluation score of 4.40 compared with 2.53 for article-based RAG ($p = 0.0039$). These empirical findings confirm that robust IoT support agents require dynamic knowledge governance that bridges versioned documentation, synchronized asset metadata, and live telemetry data within a unified retrieval framework.

Keywords: AI Agents; Customer Support; Internet of Things; Knowledge Management; Retrieval-Augmented Generation.

1. Introduction

The deployment of artificial intelligence as the primary engine for customer support services has accelerated significantly alongside the rapid maturity of large language models (LLMs) built on the Transformer architecture (Vaswani *et al.*, 2017). Enterprise organizations increasingly leverage conversational AI agents to reduce tier-one support overhead, shorten resolution intervals, and provide uninterrupted round-the-clock technical assistance. However, practical field deployments consistently demonstrate that the performance bottleneck of these agents rarely stems from the intrinsic reasoning or syntactic fluency of the underlying language models. Instead, the operational ceiling is dictated by whether the agent has real-time access to relevant, accurate, well-structured, and verifiable enterprise knowledge. When conversational models operate in isolation or query poorly maintained data repositories, they generate responses that are either factually inaccurate, outdated, or dangerously ungrounded in current operational realities. Theoretically, this deployment challenge reflects a classic problem in organizational knowledge management: the systematic externalization and conversion of tacit knowledge held within the minds of experienced technical specialists into structured explicit knowledge that automated systems can reliably retrieve and interpret (Nonaka & Takeuchi, 1995). The established knowledge management literature underlines that the architectural quality, indexing structure, updating frequency, and overall accessibility of enterprise assets directly determine the effectiveness of downstream automated information systems (Alavi & Leidner, 2001; Dalkir, 2017). Within the paradigm of conversational AI, this challenge becomes operationally acute because generative text synthesis depends strictly on the context injected into the prompt at inference time. Retrieval-Augmented Generation (RAG) has emerged as the leading standard for connecting parametric language models to external, non-parametric knowledge repositories (Lewis *et al.*, 2020). Nevertheless, the factual fidelity of RAG frameworks remains fundamentally bounded by the completeness, topical precision, and temporal freshness of the retrieved data passages (Gao *et al.*, 2024). This knowledge dependency becomes particularly vulnerable in complex technical environments such as the Internet of Things (IoT). Unlike general enterprise domains where static documents suffice, IoT customer support involves physical assets operating under continuous environmental fluctuations. Support inquiries in connected ecosystems rarely ask solely for static setup procedures; users routinely report device offline statuses, intermittent connectivity drops, anomalous sensor readings, battery depletion, or failed over-the-air updates. In such operational contexts, enterprise knowledge is inherently fragmented across multiple layers. Static standard operating procedures (SOPs) and troubleshooting manuals describe nominal device behavior, but they are entirely blind to current device health, firmware versions, active MQTT protocol messages, or localized network conditions. As Dzenuska and Rudzajs (2025) note, feeding generative agents with scattered, rarely updated, and purely static knowledge bases inevitably causes customer support agents to generate generic, incomplete, or misleading troubleshooting guidance.

Despite highlighting these challenges, the preliminary empirical study conducted by Dzenuska and Rudzajs (2025) left critical methodological and architectural limitations unaddressed. Their investigation was restricted to an evaluation of ten static documentation articles and twenty test questions derived from a single corporate setting, which significantly limits the statistical generalizability of their findings across diverse operational contexts. Methodologically, their evaluation was entirely qualitative and lacked standardized automated text-generation benchmarks, such as BERTScore (Zhang *et al.*, 2020) or Cosine Similarity, as well as randomized, blinded human assessment protocols to safeguard against subjective evaluation bias. Most fundamentally, their architecture relied strictly on static, document-based knowledge articles, completely ignoring the dynamic operational telemetry, event logs, and synchronized device metadata that characterize modern IoT ecosystems. To resolve these empirical and architectural gaps, this study presents a hybrid knowledge management framework tailored specifically for IoT customer support AI agents. The proposed approach systematically unifies static knowledge repositories—including procedural manuals, SOPs, and technical FAQs—with semi-dynamic device metadata records and streaming operational telemetry, such as real-time signal strength indicators (RSSI), battery percentages, error frequencies, and MQTT logs, within an integrated RAG pipeline. Through a controlled pilot experimental design comprising thirty representative support questions evaluated under four progressive tiers of knowledge integration, this study investigates how different levels of data contextualization affect response fidelity across informational, status-oriented, and diagnostic inquiries. Ultimately, this research aims to bridge the gap between organizational knowledge governance and real-time telemetry ingestion, providing both theoretical grounding and actionable architectural principles for building production-grade IoT support systems.

2. Related Work

This study is conceptually grounded in organizational knowledge management theory, particularly the foundational paradigm of tacit versus explicit knowledge formulated by Nonaka and Takeuchi (1995). In technical customer support environments, explicit knowledge codified in manuals, standard operating procedures, and technical documentation forms the primary substrate for automated retrieval systems. As Dalkir (2017) emphasizes, comprehensive knowledge management frameworks must facilitate effective capture, structured indexing, rapid retrieval, and practical operational application. Furthermore, Alavi and Leidner (2001) argue that systematically governed and accessible knowledge assets directly enhance overall system efficacy—a principle that applies directly to the contextual repositories underpinning conversational agents. Within conversational AI research, empirical findings confirm that disciplined knowledge governance is just as critical as model parameter scale for sustaining response reliability, minimizing hallucinations, and grounding conversational agents in enterprise realities (Laskar *et al.*, 2025; Salemi & Zamani, 2025; Shuster *et al.*, 2021). Retrieval-Augmented Generation has become the predominant paradigm for anchoring generative language models to external, non-parametric knowledge sources, thereby mitigating generative hallucinations in knowledge-intensive tasks (Lewis *et al.*, 2020; Shuster *et al.*, 2021). In a comprehensive survey, Gao *et al.* (2024) identify retrieval quality and contextual relevance as the principal bottlenecks dictating end-to-end generation fidelity. Expanding beyond standard vector search, Edge *et al.* (2024) demonstrate that structuring textual knowledge into graph-based relational abstractions substantially improves answer synthesis for complex, query-focused summarization tasks. To evaluate generated responses objectively, contemporary frameworks employ both lexical and semantic benchmarks: Cosine Similarity measures broad vector-space proximity, whereas BERTScore evaluates contextual token-level semantic equivalence using pre-trained language representations (Zhang *et al.*, 2020). Complementing reference-based metrics, automated assessment frameworks such as RAGAS evaluate dimensions of faithfulness, answer relevance, and context utilization without requiring manual ground-truth annotations (Es *et al.*, 2024). Despite these advancements in document-based retrieval, the intersection of real-time operational telemetry and retrieval-augmented knowledge architectures remains largely underexplored. The vast majority of RAG literature assumes static, narrative-driven text corpora as the sole retrieval target, largely overlooking dynamic operational data such as sensor readings, device status logs, or streaming MQTT messages. As Farahani *et al.* (2018) emphasize, IoT ecosystems generate heterogeneous, time-sensitive data streams that static technical documentation cannot represent. In IoT technical support, hardware health, active firmware versions, and real-time connectivity fluctuations directly determine troubleshooting outcomes. Consequently, relying exclusively on static documentation creates an operational disconnect, leaving conversational agents unable to verify live physical states or diagnose transient hardware anomalies.

3. Methodology

3.1 Research Design

This study adopts a mixed-method pilot research design that integrates four methodological components: a systematic literature review, a comparative method analysis, an experimental evaluation, and a conceptual exploration. This integrated framework was selected because resolving customer support challenges in complex technical environments extends beyond quantitative model benchmarking; it requires a systematic understanding of how enterprise knowledge assets are organized, maintained, and operationalized for automated reasoning. The literature review established the conceptual intersection between knowledge management principles, retrieval-augmented architectures, and conversational agents. The comparative analysis positioned the multi-tier hybrid architecture against standard document-based retrieval pipelines. The experimental evaluation quantitatively measured AI-agent answer fidelity across controlled contextual scenarios, while the conceptual exploration synthesized these empirical findings into an operational knowledge governance framework for IoT enterprises. Designed as a pilot experiment, the primary objective is not broad statistical generalization, but an empirical assessment of whether incremental knowledge integration tiers yield measurable improvements in response quality across distinct support query categories.

3.2 RAG System Architecture

The architectural pipeline evaluated in this study extends the standard Retrieval-Augmented Generation framework formulated by Lewis *et al.* (2020). In a conventional RAG configuration, user inquiries trigger a dense or sparse retriever that identifies top-k relevant textual passages from an external corpus. The user query and retrieved passages are subsequently formatted into an augmented context prompt and passed to a generative language model for response synthesis. In the baseline article-based RAG configuration, this external corpus is restricted entirely to static textual documentation, such as standard operating procedures,

troubleshooting manuals, hardware specifications, and technical FAQs. This study expands this baseline architecture by incorporating two dynamic contextual layers: structured device metadata records and real-time operational telemetry streams. Device metadata records provide semi-dynamic operational context, including unique device identifiers, installed firmware builds, installation timelines, asset ownership, and localized maintenance logs. The telemetry pipeline ingests high-velocity operational signals, encompassing network connectivity states, signal attenuation metrics, live sensor readings, battery charge percentages, MQTT protocol logs, recent error event codes, and localized device timestamps. By dynamically merging these disparate data layers into the prompt assembly stage, the hybrid architecture enables conversational agents to address procedural queries as well as live status checks and fault diagnostic inquiries that require real-time empirical verification.

3.3 Experimental Procedure

The experimental evaluation was conducted over an IoT technical support knowledge base combining static documentation, structured device metadata, and operational telemetry. Because this study represents an exploratory pilot evaluation, portions of the device metadata and telemetry data were synthetically generated to model realistic enterprise IoT support scenarios. Thirty test questions were curated and classified into three functional categories: informational questions addressing procedural steps or general product specifications; status questions querying current operational states; and diagnostic questions requesting root-cause identification and actionable corrective workflows. Simulated telemetry values were generated by defining parameter ranges calibrated against published IoT hardware specifications and empirical deployment literature. Signal strength was modeled between −50 dBm and −90 dBm to reflect typical indoor Wi-Fi and LoRa propagation environments; battery depletion followed standard lithium-ion discharge profiles (0–100%); and error event frequencies were bounded between 0 and 5 events per hour based on common IoT gateway failure profiles documented in manufacturer technical reports. Telemetry parameters were systematically varied across simulated devices to introduce diverse failure modes into diagnostic scenarios. Each of the thirty questions was evaluated across four progressive knowledge integration tiers: Scenario 1 (S1) without RAG, providing only the raw user prompt to the model; Scenario 2 (S2) article-based RAG, augmenting the prompt with retrieved static documentation; Scenario 3 (S3) document-based RAG supplemented with structured device metadata; and Scenario 4 (S4) full hybrid integration combining static documentation, device metadata, and live operational telemetry streams. For every scenario, the model generated an answer that was benchmarked against an authoritative ground-truth reference response, allowing precise observation of how incremental knowledge layers influence generation fidelity across query categories.

3.4 Automatic Evaluation Metrics

Automated response quality was evaluated using two complementary linguistic metrics. BERTScore (F1) was employed to measure deep semantic similarity between model-generated responses and reference answers using contextual embeddings derived from pre-trained transformer representations (Devlin *et al.*, 2019; Zhang *et al.*, 2020). In parallel, Cosine Similarity was calculated over dense sentence embeddings to evaluate broader vector-space proximity. BERTScore captures nuanced semantic equivalence that simple surface-level token overlap overlooks, whereas Cosine Similarity provides a stable baseline for overall topical alignment. Because the experimental design involved paired responses generated from identical test queries across distinct integration tiers, statistical significance was determined using the Wilcoxon signed-rank test. This non-parametric paired test was selected due to the pilot sample size and the non-normal distribution of evaluation scores.

3.5 Human Evaluation Procedure

To validate the clinical relevance of the automated metrics, a randomized, double-blind human evaluation was conducted. Ten representative query pairs comparing article-based RAG (S2) against the full hybrid architecture (S4) were selected using stratified sampling across query categories. Each evaluated response was anonymized, randomized, and stripped of scenario metadata to eliminate evaluator bias. The qualitative assessment was performed by a single domain researcher across three core evaluation criteria: relevance, assessing whether the response directly addresses the user inquiry; correctness, evaluating factual precision and technical validity against ground-truth conditions; and usefulness, measuring the practical actionability of the response from an end-user support perspective. Each criterion was scored on a 5-point Likert scale ranging from 1 (very poor) to 5 (very good). Following scoring, blind identifiers were decoded using a secure key. While utilizing a single evaluator introduces potential subjective variance, this protocol provides a necessary qualitative check to ensure that statistical improvements captured by automated metrics reflect meaningful utility in real-world customer interactions.

2.6 Conceptual Exploration for the IoT Domain

In the final phase of the methodology, empirical findings were synthesized into an architectural blueprint designed for production enterprise environments. This conceptual exploration analyzed the operational governance required to ingest heterogeneous IoT data streams into generative agent pipelines without compromising system latency or factual consistency. The resulting framework establishes a three-tier knowledge hierarchy: static foundational knowledge, semi-dynamic device records, and streaming operational telemetry. Static documentation requires rigorous editorial curation, semantic chunking, and metadata tagging; semi-dynamic device records necessitate continuous synchronization with enterprise asset management databases; and high-velocity telemetry requires edge filtering and event aggregation before injection into the retrieval context. This layered abstraction provides the theoretical and operational foundation for interpreting the subsequent experimental results.

4. Results and Discussion

4.1 Results

Table 1 presents the mean Cosine Similarity and BERTScore across each question category and knowledge integration scenario. The quantitative trajectory confirms the primary research hypothesis: response alignment improves systematically as contextual sources become more operationally complete. However, the magnitude of performance gains varies substantially across question categories.

Table 1. Average Automatic Scores by Question Type and Scenario (n = 30)

Question type (metric)	S1	S2	S3	S4
Informational (Cosine)	0.553	0.750	0.755	0.748
Informational (BERTScore)	0.691	0.815	0.819	0.815
Status (Cosine)	0.612	0.704	0.799	0.839
Status (BERTScore)	0.704	0.701	0.752	0.760
Diagnostic (Cosine)	0.728	0.662	0.783	0.830
Diagnostic (BERTScore)	0.657	0.701	0.704	0.704

The comparative trajectories across scenarios are illustrated in Figure 1, highlighting the distinct response behaviors observed between informational and operational query types.

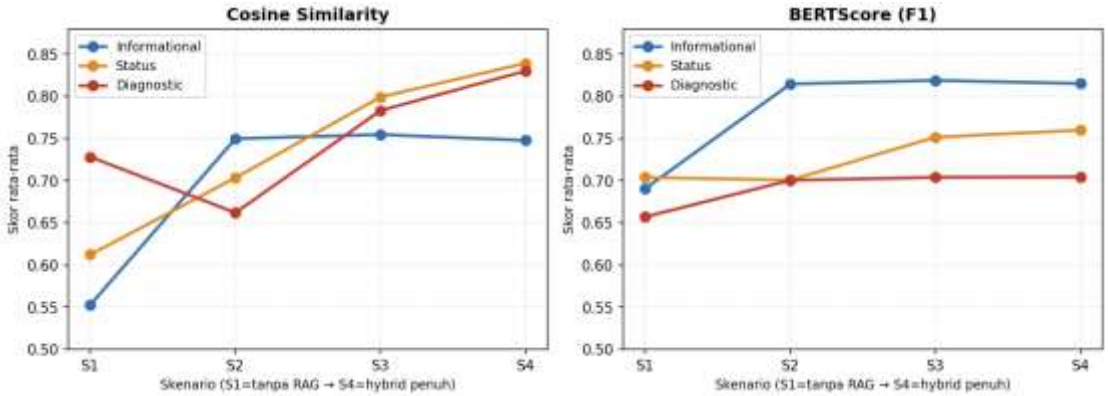

Figure 1. Trends in Cosine Similarity and BERTScore Across Scenarios, Separated by Question Type (n = 30)

For informational inquiries, metric scores rose sharply from S1 to S2, with Cosine Similarity increasing from 0.553 to 0.750 (p = 0.027), before plateauing across S3 and S4, where the change between S2 and S4 was negligible (−0.002, p = 0.750). BERTScore exhibited an identical trend, rising from 0.691 in S1 to 0.815 in S2 and remaining virtually unchanged through S4 (0.815). This plateau indicates that curated static documentation is fully sufficient for resolving procedural or conceptual questions. Augmenting the prompt with device metadata records or live telemetry streams yields no measurable benefit because procedural guidelines do not depend on real-time hardware states. Conversely, for status and diagnostic inquiries, the full hybrid configuration (S4) significantly outperformed baseline article-based RAG (S2). In status questions, Cosine Similarity improved by +0.136 (p = 0.002) and BERTScore increased by +0.059 (p = 0.004) between S2 and S4. In diagnostic questions, Cosine Similarity showed a marked gain of +0.168 (p = 0.004), although BERTScore remained essentially flat (+0.004, p = 0.846). These empirical trends demonstrate that inquiries requiring physical operational verification cannot be adequately resolved through static documentation alone.

The randomized blind human evaluation of ten representative query pairs strongly corroborates the automated linguistic metrics. Table 2 details the comparative performance across evaluation criteria, and Figure 2 presents the dimensional profile visually.

Table 2. Average Human Evaluation Scores Across Operational Criteria (Scale 1–5)

Aspect	Article-based RAG (S2)	Hybrid (S4)
Relevance	2.4	4.5
Correctness	2.7	4.6
Usefulness	2.5	4.1
Overall average	2.53	4.40

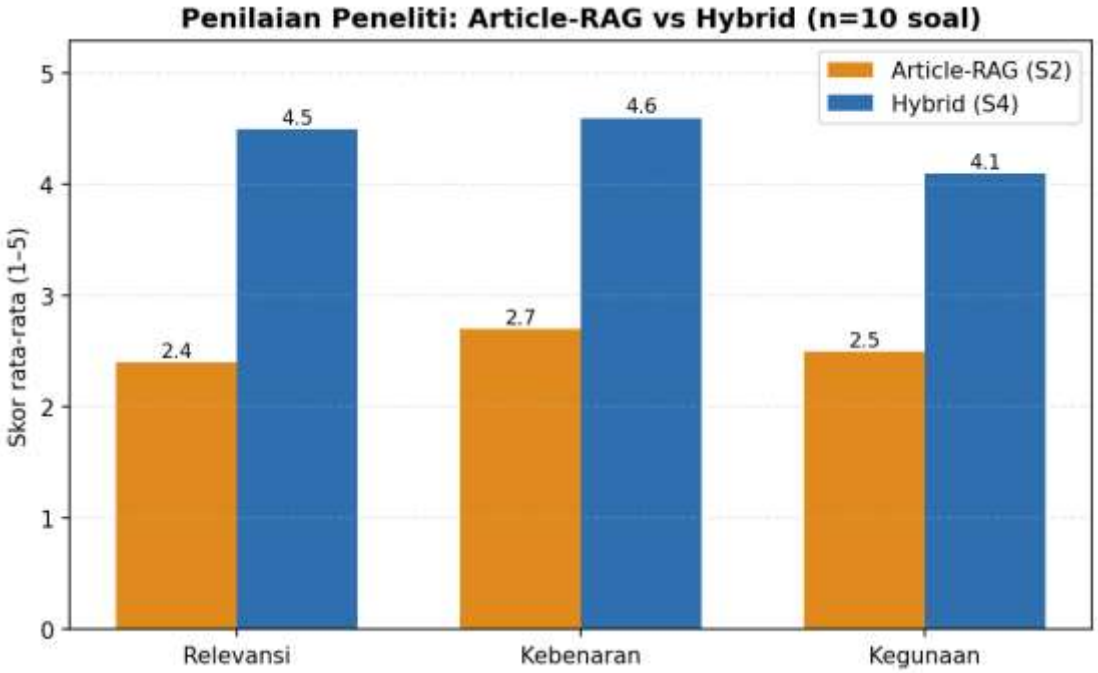

Figure 2. Human Evaluation Comparison Between Article-Based RAG (S2) and Hybrid Knowledge (S4) by Aspect

The hybrid architecture achieved an overall mean score of 4.40 out of 5.0, compared with 2.53 for article-based RAG. This difference was statistically significant under a paired Wilcoxon signed-rank test ($p = 0.0039$). At the individual question level, hybrid responses outperformed article-based RAG in nine of the ten evaluated pairs, achieved an identical score on one pair, and were never rated lower. The most pronounced qualitative gain occurred in factual correctness (rising from 2.7 to 4.6), directly attributable to the presence of real-time telemetry and synchronized device records that provided empirical verification absent from static technical manuals.

4.2 Discussion

The central empirical finding of this investigation is that the operational utility of knowledge integration in conversational AI is governed by query intent rather than applying uniformly across all user interactions. For procedural informational inquiries, static technical documentation provides complete coverage, rendering live telemetry ingestion redundant. In sharp contrast, status and diagnostic interactions inherently depend on the real-time operational state of physical hardware. Operating without telemetry confines conversational agents to generic troubleshooting checklists, whereas ingesting live telemetry enables agents to isolate device-specific failure states and formulate precise, actionable remediation workflows. A notable methodological pattern requiring critical interpretation is the divergence observed between the two automated evaluation metrics in diagnostic scenarios: Cosine Similarity increased substantially (+0.168 between S2 and S4), whereas BERTScore exhibited negligible movement (+0.004). This metric divergence stems from differences in how each metric evaluates text. BERTScore evaluates deep contextual embeddings and is comparatively insensitive to localized factual insertions when the overarching narrative framing of the troubleshooting response remains constant. In contrast, Cosine Similarity responds directly to dense vector shifts driven by the inclusion of exact, domain-specific terminology, such as discrete firmware build strings, localized error event codes, signal decibel readings, and operational status values. Consequently, automated benchmarking of technical support agents

should not depend on a single evaluation metric. These empirical findings substantiate and expand upon existing knowledge management and conversational AI literature. In alignment with Laskar *et al*. (2025), the findings demonstrate that disciplined knowledge base structuring directly dictates conversational agent performance. The results also reinforce the conclusions of Gao *et al*. (2024), who identify contextual retrieval quality as the defining bottleneck in RAG implementations. In IoT environments, however, retrieval quality extends beyond text-matching relevancy to encompass the temporal freshness and validity of operational data. Enterprise knowledge architectures for connected devices must therefore evolve beyond static document indexes to incorporate automated data pipelines that harvest live telemetry streams. Several methodological constraints qualify the conclusions of this pilot investigation. Primarily, the experimental dataset was restricted to thirty questions evaluated over a single knowledge base, and portions of the device metadata and operational telemetry were synthetically modeled, bounding broad empirical generalization. Furthermore, the qualitative evaluation was conducted by a single domain researcher, introducing potential scoring subjectivity. Additionally, the current experimental matrix did not include fine-grained ablation isolating the independent contributions of device records versus telemetry streams within Scenario 4, nor did it evaluate live end-to-end integration with production MQTT message brokers, enterprise asset databases, or active enterprise support queues. Future investigations should expand the benchmark corpus, incorporate multi-evaluator panels, integrate live telemetry streams, and conduct systematic component ablation across diverse language model backbones and vector retrieval algorithms. From an enterprise systems standpoint, operationalizing real-time telemetry within production RAG pipelines introduces critical engineering trade-offs, including retrieval latency overhead, data freshness synchronization to prevent stale telemetry from misleading reasoning agents, and strict role-based access control over sensitive device logs. Addressing these operational constraints will be essential for translating hybrid retrieval architectures into production-grade IoT support systems.

5. Conclusion and Recommendations

This pilot investigation demonstrates that the response fidelity of conversational AI agents in IoT customer support is governed by the structural alignment between user query intent and tiered knowledge integration, rather than solely by the generative capacity of the underlying language model. While curated static documentation adequately addresses procedural and informational inquiries, resolving operational status checks and hardware diagnostics requires the dynamic ingestion of device metadata and live telemetry streams. The hybrid integration framework consistently achieved superior performance across both automated linguistic benchmarks and double-blind human evaluations. Consequently, enterprise IoT organizations deploying conversational support agents must transition beyond static document repositories toward comprehensive knowledge governance architectures that systematically unify standard operating procedures, asset metadata records, and streaming telemetry pipelines within a coherent retrieval-augmented framework.

References

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, *25*(1), 107–136. https://doi.org/10.2307/3250961

Dalkir, K. (2017). *Knowledge management in theory and practice*. The MIT Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dzenuska, E., & Rudzajs, P. (2025). Toward a knowledge management method for training customer support AI agents. In *Proceedings of the BIR 2025 Workshops and Doctoral Consortium* (CEUR Workshop Proceedings, Vol. 4034, pp. 70–83). CEUR-WS.org.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). *From local to global: A graph RAG approach to query-focused summarization*. arXiv. https://doi.org/10.48550/arXiv.2404.16130

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024): System Demonstrations* (pp. 150–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16

Farahani, B., Firouzi, F., Chang, V., & Badaroglu, M. (2018). Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare. *Future Generation Computer Systems*, *78*(Part 2), 659–676. https://doi.org/10.1016/j.future.2017.04.036

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

Laskar, M. T. R., Tremblay, J. B., Fu, X.-Y., Chen, C., & Bhushan, S., TN. (2025). *AI knowledge assist: An automated approach for the creation of knowledge bases for conversational AI agents*. arXiv. https://doi.org/10.48550/arXiv.2510.08149

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems (NeurIPS 2020)* (Vol. 33, pp. 9459–9474). Curran Associates, Inc.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

Salemi, A., & Zamani, H. (2025). HELMET: How to evaluate long-context language models effectively and thoroughly. In *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*. OpenReview.net.

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems (NeurIPS 2017)* (Vol. 30, pp. 5998–6008). Curran Associates, Inc.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net.