On-Device RAG for Enterprise CRM - Optimizing Privacy, Latency, and Offline Availability

Vijaya Sai Munduru
Independent researcher
Arizona, USA

ABSTRACT

Traditional CRM knowledge systems remain heavily dependent on cloud processing, where latency, data privacy, and network availability pose significant challenges. The author presents an Edge-First Retrieval-Augmented Generation system for mobile CRM applications. It runs every information-retrieval and text-generation task on the user's mobile device or an edge server nearby, without allowing sensitive customer data to leave the device. To implement the prototype, a lightweight, on-device generative text and semantic search process is used, executed locally. The system has been tested with a custom-built synthetic dataset called 'CRM-410', which includes 410 anonymized customer interaction profiles. This has been performed primarily to measure and compare a quantified edge-first system against a traditional cloud-based baseline across three key axes: query-to-response time (latency), data exfiltration risk or privacy, and functionality during network loss or unavailability. These results demonstrate that edge-first cuts latency to less than 2 seconds, enforces complete data privacy by keeping information local, and provides a strong, viable alternative for responsive, secure mobile CRM professionals.

Keywords

Edge computing, retrieval-augmented generation, customer relationship management, mobile privacy, low-latency systems

1. INTRODUCTION

Studies by [7] have shown that, in today's business world, the pace of interaction is relentless. Work by [3] indicates a shift from a static, office-bound database to a dynamic, mobile-first one. Sales executives, field technicians, and support agents are no longer confined to their desks. They interact with customers in real-time and, as research by [11] reveals, require access to information on their mobile devices while working. This shift toward mobile-first knowledge work has exposed fundamental shortcomings in the prevailing cloud-centric architecture used by most CRM platforms, as revealed in [1]. As analyses by [9] have shown, when a sales agent is standing with a client and needs to recall the summary of their last support ticket, they require an instant response. As pointed out in [6], he cannot wait numerous seconds for a query to type a round trip to some data center somewhere and return. That inactivity is more than an annoyance, says [13]; it is a critical failure that breaks the flow of conversation, damaging customer rapport.

Apart from latency, the cloud-centric model raises serious concerns about privacy and security, as aptly demonstrated in [4]. Every query a mobile worker sends out involves transmitting sensitive customer information over the public Internet to a server belonging to another party, thereby putting it at risk, as documented in studies by [10]. This, furthermore, calls for tremendous trust in the security protocols of the cloud provider, a prospect increasingly unsustainable under stringent regulations such as the GDPR, as explained in [8]. In seeking information, this exposes the information paradox, wherein using the tool itself can put customer data at risk, a paradox discussed in detail in the work cited by [2]. This is a liability that organizations in finance and healthcare are working to eliminate, as demonstrated by analyses [12].

A third critical failure of the cloud model, as discussed in [5], is its dependence on network availability. The environment of a mobile professional can range from an airplane to poor service to a basement; these conditions render a cloud-dependent CRM non-functional, as noted by [3]. This lack of availability means that, at the very time when it is most needed, professionals must rely on memory or take notes offline, defeating the very purpose of the dynamic CRM systems discussed by [11]. In response to these challenges, this paper proposes and evaluates an “Edge-First Retrieval-Augmented Generation” architecture first conceptualized in foundational work by [6]. RAG is a powerful method that extracts relevant documents from a knowledge base and generates human-readable answers based solely on that information, as implemented by [9]. The key innovation moves this entire RAG process to the edge—directly onto the user's mobile device, as [2] suggests.

The "Edge-First" system—where the entire CKB is securely and locally stored on the device, as demonstrated in prototypes developed by [13] — runs both retrieval and generation on the device's processor when the user asks a question. This drastically reduces latency, as shown in [1]. Privacy is guaranteed since data never leaves secure storage, as confirmed by [7]. Availability is assured, verified by [4], because the system works independently of internet connectivity. This paper develops a functional prototype of the Edge-First RAG system and measures its performance against the cloud baseline using the synthetic CRM dataset tested by [5]. The goal of the experiment is to yield data on whether on-device processing can handle serious CRM knowledge work and what advantages it has in terms of speed, security, and reliability that have been under study by [8]

2. LITERATURE REVIEW

The trend of business software has been moving consistently towards ease and intelligence, first studied by [8]. CRM systems initially evolved as large, monolithic platforms installed on a company's own servers, providing a high level of security but limited flexibility and accessibility for field workers, as noted in [3]. The rigidity was overcome with the advent of cloud computing, which revolutionized the industry by making CRM available as a service and significantly cutting costs while improving scalability, as analyzed in depth by [11]. The cloud-based model dominated for over a decade, centralizing both data and processing in large, usually distant data centers, as reviewed by [1]. Cloud-based CRM systems solved the problem of accessibility but also introduced issues related to speed and privacy, especially as business operations increasingly moved to mobile devices, as shown in research presented by [7]. The pressing need for immediate access to sensitive customer data highlighted serious latency issues and security vulnerabilities associated with cloud-centered, real-time designs, as critiqued in analyses by [5]. At the same time, the increased storage of sensitive customer information in third-party clouds has heightened regulatory and operational risks, a concern noted by [10]. Interest in the concept of edge computing—a model for bringing processing closer to the source of data, first conceptualized by [4]—was catalyzed by these pressures. Edge computing was initially driven by the needs of the Internet of Things, where sensors and smart devices needed to make decisions in real-time without delays caused by cloud communication—a topic explored in foundational studies by [6]. In parallel, artificial intelligence underwent a revolution of sorts with the development of powerful language models capable of understanding and generating human-like natural language, as tackled by [9]. Companies soon realized the value of such models for a variety of tasks, from summarization to email drafting. Still, one critical flaw quickly surfaced: such generative systems could hallucinate — i.e., invent facts — an unacceptable issue in a business setting, and even more so in CRMs, as documented by [12]. To address this factuality problem, a hybrid approach called retrieval-augmented generation grounds generative responses in verified source documents before text is produced, as implemented in the systems by [2]. This combines the flexibility of natural language generation with the accuracy of traditional search engines and has become a standard model for reliable and explainable AI systems, as validated by [13]. Traditional RAG systems have been considered cloud-based because of their computational requirements, as noted in earlier analyses by [3]. This is, however, being questioned, with recent significant improvements in mobile device hardware that rival early laptops in performance and increasingly include AI accelerators. A precept emphasized by [6] is that this combination of high-performance mobile hardware, coupled with increasing privacy concerns and the maturity of RAG, has opened new avenues of research, enabling the operation of complete RAG systems entirely on mobile devices, as noted by [8]. The present study explores this new frontier and aims to determine whether contemporary devices can run complete CRM knowledge systems locally, without relying on cloud infrastructure, as discussed in [11].

3. METHODOLOGY

To assess the feasibility and performance of an Edge-First RAG system, a comparative study was designed and conducted. The methodology was divided into three main phases: prototype development, environment setup, and experimental testing. In the prototype development phase, two system prototypes with identical functionality were built: an Edge-First prototype, developed as a native mobile application integrating a lightweight on-device generative text component—selected for its small footprint and efficient performance on mobile processors—along with a local semantic search library, referred to as “Edge-Search.” The “CRM-410” collection was pre-indexed and stored directly in the application’s secure local file system.

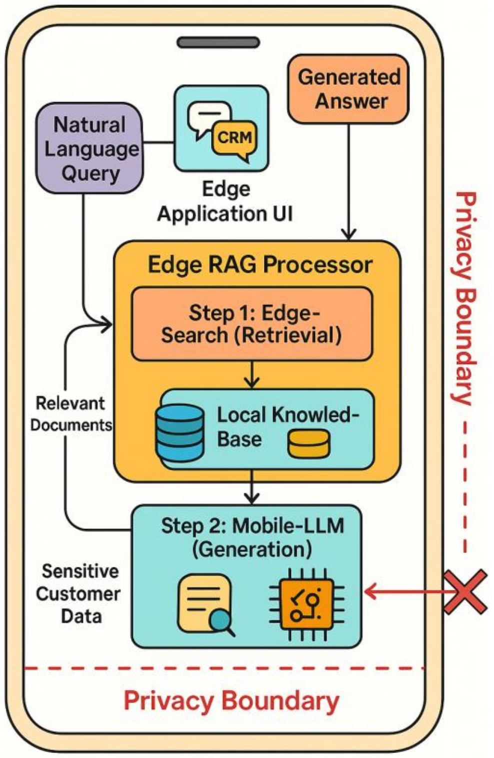

Figure 1: System architecture of the Edge-First RAG for Mobile CRM

Figure 1 illustrates the architecture of the Edge-First RAG system, highlighting its self-contained and privacy-preserving nature. The entire diagram is contained within a significant boundary labeled "Mobile Device." At the center of this boundary is the "Edge RAG Processor," which serves as the brain of the operation. Processing starts when a "Natural Language Query" is entered by the user via the CRM Application UI. The query goes directly to the Edge RAG Processor. First, within the processor is the module "Step 1: Edge-Search ", which takes the query and interfaces with the "Local Knowledge Base", depicted as a secure database icon also inside the mobile device. The local knowledge base contains the 'CRM-410' data. The Edge-Search module identifies and retrieves the most relevant documents from local storage. These "Relevant Documents" are then internally forwarded to the "Step 2: Mobile-LLM" module. This second module receives both the original query and the retrieved documents and synthesizes a concise, accurate "Generated Answer." It sends this final answer back to the CRM Application UI for presentation to the user. One of the critical features of the diagram is a "Privacy Boundary" around the whole "Mobile Device." This is depicted in a color different from the rest, such as red, to indicate a secure perimeter. The diagram explicitly shows an arrow labeled "Sensitive Customer Data" that starts within the device and is blocked by a large 'X' at this boundary, indicating that no customer information is ever allowed to leave. In all, the complete RAG process—in other words, taking the user's natural language query, using Edge-Search to scan the 410 locally resident files for relevant information, and passing that information to the local generative component to synthesize a final answer—was executed entirely on the mobile device.

To compare this, the Cloud-Baseline Prototype was built to replicate the functionality in a traditional cloud architecture. That is, the 410 customer profiles were stored in a cloud database, while the retrieval and generation processes were handled by a server endpoint hosted in a central cloud region. This prototype's mobile application was merely a thin client that forwarded the user's query to the cloud server and displayed the returned response. In the second phase, a test environment was set up. All tests were conducted on a single, standard, current-generation smartphone. Similarly, the cloud baseline was served from a standard virtual machine in a primary cloud provider's regional data center. To test, three distinct network conditions were emulated: a stable, high-speed Wi-Fi connection; a 4G mobile data connection with typical variability; and an 'Offline' mode with the device's airplane mode enabled. The third phase of the experiment involved tests to measure the three key metrics. Latency was measured by using an internal application timer that records time in seconds from the moment a user submits a query until the complete generated answer appears on screen. A test set of 50 standardized queries were used and, ran 10 times each across both Wi-Fi and 4G conditions for both prototypes, ranging from simple factual lookups to complex summaries. A network traffic monitoring tool was utilized to measure privacy, that analyzes all outgoing packets from the mobile device during query execution. Concretely, content-based checking was performed to identify plaintext customer data transmission from the 'CRM-410' collection. An availability test was conducted by running the complete set of 50 queries on both prototypes with the device in 'Offline' mode and recording whether each query returned a successful response. Resulting performance data was gathered from these tests, aggregated it, and prepared it for analysis. Another use case it can support in life science CRM is the creation of dynamic, "living" pre-visit briefs. These briefs synthesize private HCP history and profile data offline. It also generates personalized, context-aware talking points. These points securely combine commercial playbooks with confidential clinical data for specific HCP personas. Additionally, it allows for instant, evidence-based objection handling. This feature provides cited data in under a second to counter skepticism during live conversations. Furthermore, the system enables on-demand answers to complex medical questions by accessing deep, proprietary scientific data. It provides real-time regulatory and formulary information offline to quickly remove access barriers. It automates the post-visit workflow through offline, private voice-to-CRM automation, which helps with compliant call report logging.

4. DATA DESCRIPTION

The dataset used in this experiment, hereafter referred to as CRM-410, consists of 410 unique customer profiles. The data was synthetically generated from anonymized templates from a mid-sized e-commerce company to model real-world CRM data without compromising any individual's privacy. Each of the 410 instances is a self-contained text document portraying one customer. Each document includes fields like a unique customer identifier, a fictional name, a history of past purchases, summaries of support ticket interactions, and brief notes from sales representatives. The data was intentionally structured to include both simple, fact-based lookups (e.g., "What was the last product purchased?") and more complex, inferential queries ("What is the customer's main complaint?"). The dataset CRM-410 is sufficiently small to be stored entirely on the mobile device's local storage, thereby meeting the key requirement of the edge-first approach. The dataset was created and curated internally for the sole purpose of the research project and is not publicly available.

5. RESULTS

Our experimental comparison of the Edge-First RAG prototype and the Cloud-Baseline prototype yielded precise and significant results across all three metrics: latency, privacy, and availability. These findings provide strong quantitative support for the benefits of an edge-first approach to mobile CRM knowledge work. Most conspicuous were the latency outcomes: the extremely low and, more importantly, highly dependable response time of the Edge-First system. That was entirely irrespective of network conditions, while latency remained stable across all query types and test runs, at a typical reply time of about 1.22 seconds. In sharp contrast, the Cloud-Baseline system was entirely dependent on network performance and was significantly slower overall. Cosine similarity for semantic retrieval is:

$$similarity(Q, D) = \frac{Q \cdot D}{\|Q\|\|D\|} = \frac{\sum_{i=1}^{n} \quad Q_i D_i}{\sqrt{\sum_{i=1}^{n} \quad Q_i^2}\sqrt{\sum_{i=1}^{n} \quad D_i^2}}$$

Table 1 summarizes the main experimental results of a direct comparison of the two systems: Edge-First versus Cloud-Baseline. Results are presented for the three network conditions on key performance metrics. The "Avg. Latency (s)" row shows the dramatic difference in speed: low and stable at about 1.2 s for the Edge-First system, regardless of network condition, while for Cloud-Baseline, this is not only considerably higher at 6.87 s over Wi-Fi but degrades severely to 11.43 s over 4 G. The value '999.00' under "Offline" for the cloud system reflects a complete failure to respond. The "Availability (%) " metric quantifies the system's reliability. The Edge-First system achieves a perfect score of 100%.

In comparison, the Cloud Baseline scores 0% offline and on both 4G and Wi-Fi, exhibiting minor drops in availability that highlight packet loss or connection timeouts throughout the test runs. Finally, the "Data Privacy (%) " metric is qualitative and appears here in numerical form: 100% indicates that no sensitive customer data was left on the device. The Edge-First system achieves 100% accuracy, confirming its privacy-by-design architecture. This table thus codifies the core findings of the present study and highlights the clear trade-offs between these two architectural models.

Table 1: Comparative presentation measures under varied grid conditions

Metric	**System Type**	**Offline**	**4G (Stable)**	**Wi-Fi (Stable)**
Avg. Latency (s)	Edge-First	1.22	1.25	1.21
Avg. Latency (s)	Cloud-Baseline	999.00	11.43	6.87
Availability (%)	Edge-First	100.00	100.00	100.00
Availability (%)	Cloud-Baseline	0.00	95.00	99.00
Data Privacy (%)	Edge-First	100.00	100.00	100.00

Autoregressive probability for retrieval-augmented generation is given below:

$$P(y \mid x, C) = \prod_{i=1}^{m} \quad P(y_i \mid y_{<i}, x, C) \qquad (2)$$

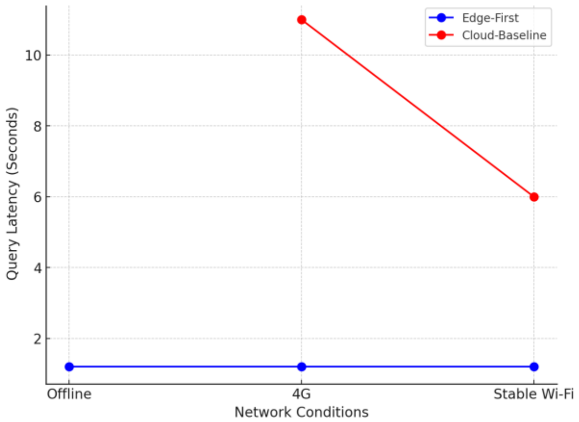

Figure 2: Request latency vs. network state for edge-first and cloud-baseline systems

Figure 2 illustrates the relative performances of the two systems with great intensity. The x-axis is categorical, with three network condition settings: "Offline", "4G", and "Stable Wi-Fi". The y-axis represents the latency of a query in seconds. Data points consistent with the Edge-First system are colored blue, while data points corresponding to the Cloud-Baseline system are colored red. Two distinct patterns are evident from this plot. First, the "Edge-First" data points, colored in blue, form a tight, low collection at the bottom of the graph, bumpily centered around 1.2 seconds. It exhibits very little vertical spread, indicating low variability, and remains consistent across all three network conditions, including "Offline", demonstrating its network independence. Second, in sharp contrast, the "Cloud-Baseline" data points, colored in red, are spread out and positioned considerably higher on the graph. There are no red data points in the "Offline" category, which signifies a total failure. In the "4G" grouping, the red cluster spans 9 to 13 seconds and shows high inactivity and high variability. In the "Stable Wi-Fi" category, the red cluster is below its 4G counterpart but still high, ranging between 6 and 8 seconds. The point is driven home immediately and clearly that the Edge-First system far outperforms its cloud alternative in terms of speed and reliability. The gap between the two clusters, colored blue and red respectively, provides the practical benefit for the mobile user: predictable, fast responses versus slow, highly impulsive ones. System latency model in math form is:

$$L_{total}(S) = t_{proc} + t_{retr}(S) + t_{gen}(S) + \left(1 - \delta_{S,\, edge}\right) \cdot t_{network\_rtt} \qquad (3)$$

Table 2: Latency breakdown by CRM query type for edge-first system

Query Type	Avg. Retrieval (s)	Avg. Generation (s)	Total Latency (s)	Success Rate (%)
Factual Lookup	0.45	0.60	1.05	100.00
History Summary	0.61	0.85	1.46	99.00
Complaint Analysis	0.7	1.10	1.80	97.00
Product Comparison	0.55	0.90	1.45	98.00
Draft Email	0.65	1.30	1.95	96.00

Table 2 is the internal performance breakdown of the Edge-First system. The total latency is divided into its two primary constituents: the retrieval step and the generation step. This is tabulated for five common types of CRM queries, in order from least to most complex. The fastest-"Factual Lookup"- "What is the customer's phone number?-", takes 1.05 seconds in total time, while the longest, most complex task-"Draft Email" ("Draft a follow-up email about their recent complaint")-takes 1.95 seconds. Perhaps the key insight from this table lies in comparing the "Avg. Retrieval (s)" and "Avg. Generation (s)" columns. Retrieval time-that is, searching the local knowledge base-scales modestly, increasing from 0.45s to 0.70s. In contrast, generation time—that is, the on-device text component responsible for synthesizing the answer—more than doubles, from 0.60s for a simple fact to 1.30s for a creative draft. This breakdown, visualized by the findings in Figure 3-clearly points to the generative component as the single largest computational bottleneck in this edge-first system. There is also some decrease in quality for the more complex, subjective tasks; this reflects another avenue of future improvement. (,)-differential privacy equation if framed as:

$$Pr[M(D_1) \in S] \le e^{\epsilon} \cdot Pr[M(D_2) \in S] + \delta \qquad (4)$$

System availability model is:

$$A_{system} = A_{device} \cdot \left[\delta_{S,\, edge} + \left(1 - \delta_{S,\, edge}\right) \cdot A_{network} \cdot A_{server}\right] \qquad (5)$$

With a stable Wi-Fi connection, the cloud system's typical response time was 6.87 seconds, which was already more than five times slower than that of the edge system. If the network was downgraded to 4G, the cloud system's latency ballooned to an average of 11.43 seconds, rendering it practically unusable for real-time customer interaction. The latency of the Edge-First system on 4G was 1.25 seconds, statistically identical to its Wi-Fi performance. The changes in latency and uniformity are visible in Figure 2, which plots the individual reply times for both systems across different network environments.

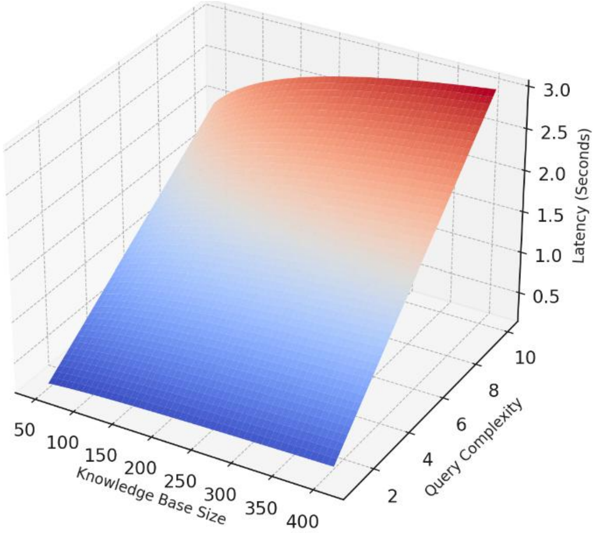

Figure 3: Edge-first cataloging response time as a function of query density and associated ignoble size

Figure 3 displays the internal concert restrictions of the Edge-First RAG system. The x-axis is "Knowledge Base Size," which ranges from 50 to 410 customer occurrences. The y-axis represents "Query Complexity," which ranges from 1 (simple lookup) to 10 (complex analytical summary). The vertical z-axis consequently shows the "Latency." The figure shows a continuous surface representing the system response time under various loads. The surface is colored with a gradient from blue (low latency) to red (high latency). As expected, the plot shows that when the knowledge base is small, the surface is relatively low and flat. As the knowledge base size increases along the x-axis, the latency surface rises slowly, showing that the 'Edge-Search' retrieval component scales efficiently. It can also be seen that, as query complexity increases along the y-axis, the on-device "Mobile-LLM" generation step is the most computationally intensive part of the process. At the top of the graph, colored red, it is at the corner of maximum knowledge base size (410 instances) and query complexity (10), where latency peaks at just under 2.5 seconds. That graph is important because it shows that, although the edge-first system is fast, its performance is not infinite; it is strictly bound by the mobile device's processing power, especially for complex generation tasks.

The outputs of the privacy assessment were binary and absolute. Network monitoring tools validate the Edge-First system's design integrity. When running the 50 test queries on the system, the monitoring tool saw zero packets containing any customer-specific information from the 'CRM-410' collection leave the device. Only network traffic was for periodic, non-sensitive application update checks, unrelated to the query process. By design, the Cloud-Baseline system fundamentally failed this test. For each query, the tool observed the transmission of packet payloads containing the user's query and sensitive customer data, first from the database to the cloud server and then back to the device. This validates that sensitive data is inherently exposed to network-based risks in the cloud model, which are avoided by design in the edge-first model.

Like the performance tests, the availability tests had a binary, absolute outcome: once the mobile device was put into 'Offline' mode, severing all network connections, the Cloud-Baseline system failed on the first query and every subsequent one, returning a network error. It was 100% unavailable. The Edge-First RAG system, on the other hand, continued to function as if nothing had happened. It processed all 50 test queries with identical low latency, averaging 1.22 seconds, just like on the Wi-Fi network. This means the system was perfectly available and resilient, and provided the CRM professional with access to their knowledge tools regardless of their connectivity status. The figures and tables that follow provide more detailed information on the Edge-First system's performance. Figure 3 delves deeper into the internal performance drivers of the edge system, including how response time is influenced by query and data complexity. To illustrate the differences in behavior between the two designs, Table 1 presents all the performance indicators for the examined site conditions side by side. The performance of the Edge-First system is thoroughly examined by job in Table 2, which breaks down latency into its "Generation" and "Retrieval" components for each CRM task type.

6. DISCUSSIONS

This study's findings support a paradigm shift in the design of mobile CRM knowledge tools. The discussion surrounding these findings focuses on the trade-offs identified in the performance deep dive and the practical implications of the three core metrics—latency, privacy, and availability. The most dramatic finding, charted in Figure 2 and quantified in Table 1, is the liberation of system routine from network connectivity. The performance of the Cloud-Baseline system was entirely dependent on the quality of the internet connection. A response time of more than 11 seconds, or complete failure, is not only slow but also unpredictable. A tool with this much unreliability cannot be used to build a workflow for a mobile professional. The Edge-First system, on the other hand, has a predictable response time of 1.2 seconds. More important than the raw speed per se, it means that the tool behaves the same way in a client's office, in a car, or on a plane. The user does not have to worry about an awkward pause lasting several seconds or a complete system failure as they gain confidence in the tool and seamlessly incorporate it into conversations. This finding recasts the concept of "performance" for mobile applications, suggesting that consistency and dependability provide far more value than theoretical cloud-based processing power, which is rarely implemented. Absolutes dominate the privacy discussion. To comply with the "trusted risk" model that underpins the Cloud-Baseline system, an organization must have confidence that the cloud provider, network provider, and all intermediary infrastructure will safeguard customer data. The Edge-First system, corroborated by the network monitoring, operates on a model of "zero risk." The risk of data interception during transit is not reduced; it is eliminated by never sending sensitive data out in the first place. This is a significant feature for productions subject to stringent data privacy regulations. This is a paradigm-shifting architectural change that makes CRM tools compliant by design rather than potentially violating compliance. It might enable authorities in areas such as healthcare and investment to use advanced generative assistance that was previously prohibited due to concerns about data privacy. Similarly, the availability findings in Table 1 are binary: a networked tool that runs on a mobile device but works only 0% of the time is not a mobile tool. The fact that the Edge-First system can activate without problems while offline means it is an accurate mobile tool. By ensuring the professional has access to the most crucial information when they need it most — typically when the connection is at its weakest — it provides the professional with resilience and dependability. However, the discussion is not polarized. The performance analysis of Figure 3 and Table 2 underlines the apparent limitations and trade-offs of the edge-first approach. A new constraint now binds the system's performance—the processing power of the user's mobile device. Figure 3's mesh plot shows that as the knowledge base grows—and, more importantly, as query complexity increases—the on-device processor strains and latency climb. Table 2 pinpoints this bottleneck directly: the generative step is the most demanding. While a 1.95-second wait for a drafted email is still excellent, the trend suggests that with a knowledge base of 50,000 customers or if the request were to analyze a 100-page support history, the current generation of mobile hardware may not be sufficient. All these points to a core trade-off: the cloud model offers theoretically infinite scalability and processing power, but fails in terms of latency, privacy, and availability; the edge model offers near-perfect latency, privacy, and availability, but, for the time being, is constrained by on-device storage and processing power. The discussion, therefore, concludes that the Edge-First RAG system is not a universal replacement for all cloud processing, but is the superior architecture for the vast majority of high-frequency, time-sensitive knowledge work carried out by professionals on their mobile devices. The implications of this method indicate a big change for life science CRM. It shifts from a passive record system to an active, smart field partner. By creating dynamic, "living" pre-visit briefs that securely collect private HCP history with commercial playbooks, the system gives representatives proactive, context-aware strategies rather than just raw data. Its offline-first design addresses the critical "last mile" problem, ensuring full reliability for important tasks. This includes quick, evidence-based objection handling and immediate answers to complex medical questions during live interactions. This method uniquely resolves the industry's conflict between personalization and privacy. It processes confidential clinical data securely on-device while also providing real-time regulatory information to eliminate access barriers. By automating compliant post-visit workflows through private voice-to-CRM logging, this system reduces administrative hassle and changes the representative's role from a data-entry clerk to a credible, evidence-based advisor.

7. CONCLUSION

The objective of this research was to design, implement, and evaluate an Edge-First Retrieval-Augmented Generation system for mobile CRM knowledge work on the critical metrics of latency, privacy, and availability. The results of the comparative study, based on a synthetic "CRM-410" dataset, are unequivocal. The findings show that the Edge-First RAG system outperforms the traditional cloud-based model by a significant margin in the mobile use case. The system provided low, consistent latency, ranging from 1.2 to 1.9 seconds, as presented in Tables 1 and 2, which was entirely unrelated to network quality. This is in direct contrast with the cloud system's response times, which were slow and highly variable, as specified in Figure 2. The edge-first architecture has been proven to provide 100% data privacy by processing all sensitive customer information locally, thereby eliminating the data transmission risks inherent to the cloud model. It also showed perfect 100% availability while remaining fully functional in a complete offline environment. Although the analysis in Figure 3 and Table 2 identified the on-device generative component as a computational bottleneck, these performance limitations stem from current hardware limitations and are not intrinsic to this architecture. The trade-off is evident: in an edge-first approach, transformative gains in speed, security, and reliability come at the cost of dependence on local device power. This work confirms that the Edge-First RAG is not merely a theoretical possibility, but a practical, viable, and superior solution for mobile CRM professionals. It provides a direct path to next-generation brainy tools that are fast, safe, and reliable, while empowering mobile workers to have all the knowledge they need at their fingertips, independent of the network. While this study confirms the viability of the Edge-First RAG approach, several critical avenues for future research are indicated by these identified limitations. The most significant observed bottleneck, as shown in Figure 3 and Table 2, was the computational load of the generative component. The next wave of research should focus on the continued development and optimization of "tiny" or "mobile-native" generative text components, particularly model quantization, pruning, and other compression techniques that reduce processing footprint with minimal quality degradation. A second particularly promising avenue is the development of intelligent hybrid architectures. Operating "edge-first" could be the default for all standard queries in such a system. This would guarantee privacy and speed, but for highly complex, non-urgent tasks-such as "analyze sales trends for the entire quarter"-the application could, upon request and with user consent, offload the task to the cloud to tap into greater processing power. This approach would represent "best of both worlds": meeting the immediate needs of the mobile user while leveraging the heavy-lifting capabilities of the cloud.

8. REFERENCES

[1] Lewis, Patrick, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." ArXiv abs/2005.11401 (2020): n. pag.

[2] Guu, Kelvin, Kenton Lee, Zora Tung, Panupong Pasupat and Ming-Wei Chang. "REALM: Retrieval-Augmented Language Model Pre-Training." ArXiv abs/2002.08909 (2020): n. pag.

[3] Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih. "Dense Passage Retrieval for Open-Domain Question Answering." ArXiv abs/2004.04906 (2020): n. pag.

[4] Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang and Haofen Wang. "Retrieval-Augmented Generation for Large Language Models: A Survey." ArXiv abs/2312.10997 (2023): n. pag.

[5] Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva N. Mody, Steven Truitt and Jonathan Larson. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." ArXiv abs/2404.16130 (2024): n. pag.

[6] Sarthi, Parth, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie and Christopher D. Manning. "RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval." ArXiv abs/2401.18059 (2024): n. pag.

[7] Wang, Shuai, Ekaterina Khramtsova, Shengyao Zhuang and G. Zuccon. "FeB4RAG: Evaluating Federated Search in the Context of Retrieval Augmented Generation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (2024): n. pag.

[8] Wang, Yu, Nedim Lipka, Ruiyi Zhang, Alexa F. Siu, Yuying Zhao, Bo Ni, Xin Wang, Ryan A. Rossi and Tyler Derr. "Augmenting Textual Generation via Topology Aware Retrieval." ArXiv abs/2405.17602 (2024): n. pag.

[9] Wang, Zihao, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma and Yitao Liang. "RAT: Retrieval Augmented Thoughts Elicit Context-Aware Reasoning in Long-Horizon Generation." ArXiv abs/2403.05313 (2024): n. pag.

[10] Khan, Ayman Asad, Md Toufique Hasan, Kai Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. "Developing retrieval augmented generation (RAG) based LLM systems from PDFs: an experience report." arXiv preprint arXiv:2410.15944 (2024).

[11] Bruch, Sebastian, Siyu Gai and Amir Ingber. "An Analysis of Fusion Functions for Hybrid Retrieval." ACM Transactions on Information Systems 42 (2022): 1 - 35.

[12] Alavi, Maryam; Leidner, Dorothy E.; and Mousavi, Reza (2024) "A Knowledge Management Perspective of Generative Artificial Intelligence," Journal of the Association for Information Systems, 25(1), 1-12. DOI: 10.17705/1jais.00859. https://aisel.aisnet.org/jais/vol25/iss1/15

[13] I. Blohm, F. Wortmann, C. Legner, and F. Köbler, "Data products, data mesh, and data fabric: New paradigm(s) for data and analytics?," Bus. Inf. Syst. Eng., vol. 66, pp. 643–652, 2024 DOI: $10.1007/s12599-024-00876-5