IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 June 2025, accepted 6 July 2025, date of publication 15 July 2025, date of current version 21 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3589241

RESEARCH ARTICLE

RAG-Driven Memory Architectures in Conversational LLMs–A Literature Review With Insights Into Emerging Agriculture Data Sharing

NUR ARIFIN AKBAR[1], (Member, IEEE), RAHOOL DEMBANI[2], BIAGIO LENZITTI[1], AND DOMENICO TEGOLO[1]

[1]Dipartimento di Matematica e Informatica, Università degli Studi di Palermo, 90123 Palermo, Italy
[2]SingularLogic, 155 61 Athens, Greece

Corresponding authors: Nur Arifin Akbar (nurarifin.akbar@unipa.it) and Domenico Tegolo (domenico.tegolo@unipa.it)

This work was supported in part by ENTRUST Project funded by European Union's Horizon Research and Innovation Program under the Marie Skłodowska-Curie Grant (EU's MSCA) under Grant 101073381.

ABSTRACT Despite significant advances in natural language processing, conversational AI systems face persistent challenges in maintaining extensive and contextually coherent dialogues, particularly regarding long-term memory management. This literature review synthesizes current approaches to memory architectures in conversational AI, examining the transition from basic dialogue agents to more sophisticated, agentic frameworks. We analyze how vector databases and Retrieval-Augmented Generation (RAG) address fundamental challenges in storing and retrieving conversational context, maintaining system responsiveness, managing user-specific data ethically, and integrating domain-specific information. Through systematic review of papers, we identify critical limitations of vector embedding in capturing extended conversational context, particularly in agentic domains requiring semantic, episodic, procedural, and emotional memory. We evaluate how RAG frameworks can augment vector databases to handle memory-intensive tasks requiring real-time updates and domain-specific knowledge integration. Furthermore, we examine alternative architectures including knowledge graphs, finite state machines, and hybrid solutions, highlighting the data quality and ethical challenges that must be addressed for scalable, reliable AI memory management. Our analysis provides a structured framework for understanding memory evolution in conversational AI, identifies gaps in current RAG solutions, proposes hybrid memory designs, and outlines future research directions emphasizing cross-domain applications in agriculture.

INDEX TERMS Artificial intelligence, conversational AI, ethical AI, hybrid memory architectures, knowledge integration, long-term memory, retrieval-augmented generation (RAG), vector databases.

I. INTRODUCTION

Conversational AI has rapidly evolved to power chatbots, digital assistants, and domain-specific dialogue systems, offering near-human fluency in many tasks [1], [2]. However, maintaining coherent multi-turn dialogues remains problematic, as systems often struggle with long-term memory and context retention [3], [4]. While advances in transformer-based models and vector databases have addressed some of these issues by enabling fast retrieval of contextually similar information [5], [6], purely embedding-centric strategies can overlook more nuanced memory types—semantic, episodic, procedural, or emotional—that are critical for agentic AI tasks like autonomous decision-making [7], [8]. As systems become increasingly personalized and domain-aware—ranging from healthcare to agriculture—new challenges emerge around data quality, ethical storage of sensitive user interactions, and

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

the integration of heterogeneous data streams [9], [10], [11], [12]. This article examines how vector databases, retrieval-augmented generation (RAG), knowledge graphs, and hybrid memory architectures collectively shape the evolution of conversational AI, with an emphasis on balancing system performance and ethical considerations. The primary goal of this literature review is to synthesize approaches to memory management in conversational AI, paying particular attention to how vector databases and Retrieval-Augmented Generation (RAG) can facilitate the transition from basic dialogue agents to more comprehensive, agentic AI frameworks [13], [14], [15]. By examining the strengths and pitfalls of vector-based retrieval, knowledge graph-based approaches, and hybrid solutions, this review clarifies how memory systems can be adapted or integrated to handle diverse types of data, from textual logs to sensor inputs [9], [16], [17], [18]. In tandem, we examine the role of data quality—accuracy, completeness, consistency, and timeliness—in ensuring the reliability of these architectures [9], [12], [19], [20].

A. THEORETICAL CONTRIBUTIONS

This paper advances the field of conversational AI memory architectures through four key theoretical contributions:

1) **Unified Memory Taxonomy**: We propose a comprehensive framework that integrates semantic, episodic, procedural, and emotional memory types, extending beyond traditional vector-based approaches to address the full spectrum of agentic AI requirements
2) **Agricultural Domain Adaptation**: We identify and formalize the unique memory challenges in agricultural data sharing, including temporal variance, multi-stakeholder access patterns, and heterogeneous data integration requirements
3) **Hybrid Architecture Design Principles**: We establish design principles for combining vector databases, knowledge graphs, and finite state machines, demonstrating how these components can be orchestrated to overcome individual limitations
4) **RAG Enhancement Framework**: We extend traditional RAG architectures to handle diverse memory types through specialized adaptation strategies, enabling more nuanced retrieval and generation capabilities.

These contributions provide both theoretical foundations and practical pathways for implementing memory-enhanced conversational AI systems in complex, real-world domains.

B. MANUSCRIPT ORGANIZATION

Figure 1 illustrates the logical flow of this review, showing how each section builds upon previous foundations to develop a comprehensive understanding of memory architectures in conversational AI.

The review proceeds through a set of significant sections which investigate: foundational concepts in conversational AI, the advent of agentic systems and their memory demands,

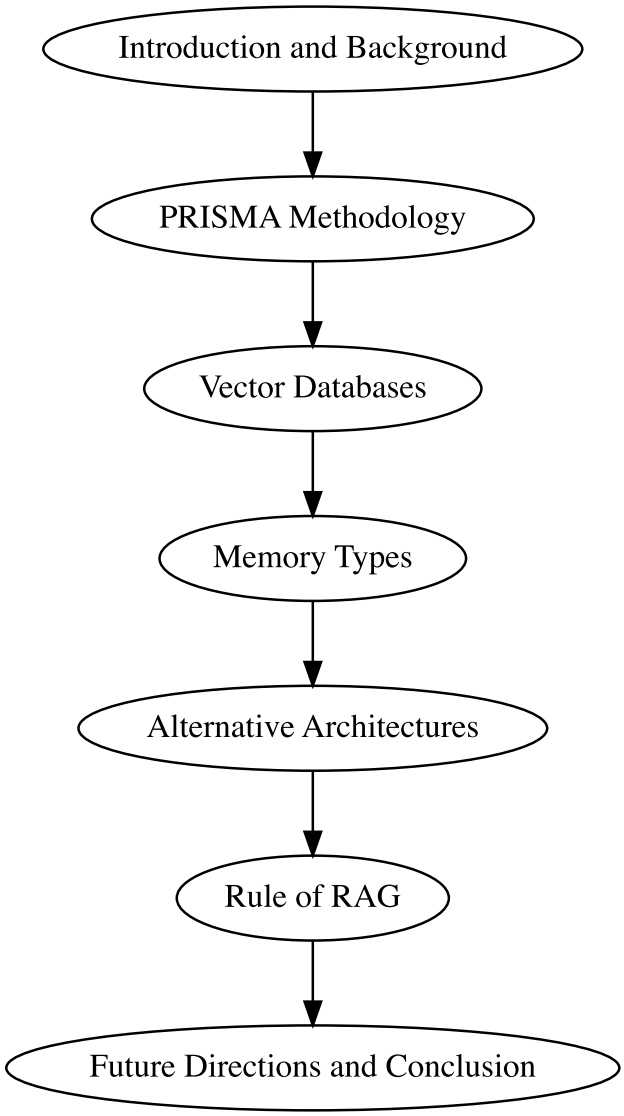

FIGURE 1. Manuscript structure showing the progression from foundational concepts through empirical validation to future research directions.

vector similarity search, data quality concerns, integrating multiple memory modalities, surveying alternatives beyond vector databases, and concluding with a discussion on RAG's future trajectory in agentic AI. Each section is anchored by a systematic PRISMA-based selection of relevant contributions(studies, articles, website,ect). PRISMA, which stands for *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*, is a widely adopted framework for conducting systematic reviews and meta-analyses in a transparent and reproducible manner. The PRISMA guidelines provide a structured approach to identifying, screening, and selecting studies, ensuring that the review process is comprehensive and unbiased. This methodology is particularly valuable in interdisciplinary fields like agricultural AI, where diverse sources of information—ranging from peer-reviewed articles to technical reports—must be synthesized to inform system design.

The PRISMA framework consists of four key stages: identification, screening, eligibility, and inclusion. During the *identification* stage, all potentially relevant studies are gathered from databases, websites, and other sources. The *screening* stage involves removing duplicates and excluding studies that do not meet predefined criteria. In the *eligibility* stage, full-text articles are assessed for relevance based on detailed inclusion criteria. Finally, the *inclusion* stage results in a curated set of studies that form the basis for analysis. By following these steps, this work ensures that only high-quality and relevant contributions are incorporated into the discussion.

The use of PRISMA underscores the rigor applied in selecting contributions for this study. It ensures that the findings are grounded in a transparent and reproducible methodology while addressing the complexity of integrating diverse data sources in agricultural AI systems, ensuring that our synthesis captures both dissemination works and cutting-edge developments [5], [6], [21]. Ultimately, this structured analysis underscores not only the current capabilities of AI memory but also the uncharted spaces demanding further research and cross-disciplinary collaboration.

C. SIGNIFICANCE OF AGRICULTURAL DATA SHARING

The agricultural sector presents unique challenges and opportunities for conversational AI systems due to its complex data ecosystem. Modern precision agriculture generates heterogeneous data streams from IoT sensors, satellite imagery, equipment telemetry, and farmer observations [11], [22]. Effective data sharing between stakeholders (farmers, agronomists, policymakers) requires AI systems capable of:

- Context-aware integration of temporal sensor data with historical records
- Secure multi-party access control for sensitive operational data
- Interpretation of unstructured farmer observations through NLP
- Reconciliation of conflicting data from disparate sources

These requirements align with the core challenges in conversational AI memory architectures, making agriculture an ideal test-bed for advanced RAG implementations. The sector's need for real-time decision support while maintaining data sovereignty requires the design of an innovative memory system [23]. These agricultural data characteristics—heterogeneous, temporal, multi-stakeholder, and mission-critical—make this domain an ideal test case for examining memory architecture challenges in conversational AI systems.

D. BACKGROUND ON CONVERSATIONAL AI AND MEMORY LIMITATIONS

While AI systems have evolved across domains from customer service to healthcare and education [1], [2], [4], [13], [24], [25], they face persistent challenges in memory management. These challenges manifest across four critical dimensions: (1) contextual continuity, where minor inconsistencies accumulate across conversations into significant coherence failures; (2) data integrity in shared repositories, where concurrent access creates risks of conflicting updates and unintended information leakage between user sessions; (3) knowledge fragmentation, as relevant information scattered across multiple embeddings or indices requires complex ''stitching'' to form coherent narratives; and (4) privacy considerations when storing personal preferences, emotional states, or domain-specific information. Vector embeddings alone struggle to address these concerns, particularly for extended multi-session dialogues. Effective conversational AI thus requires an integrated approach: structured storage architectures to prevent fragmentation, robust concurrency protocols, principled data retention policies for sensitive information, and clear boundaries between personal and global knowledge domains. These technical foundations transform fragmentary exchanges into coherent, contextually-aware conversations that build trust across interactions [3], [26], [27], [28]. Early conversational systems relied primarily on rule-based approaches with limited contextual awareness [2], [24], while today's models leverage vector databases [5], [6], [29] and knowledge graphs [30], [31], [32] to enhance memory capabilities. Despite these advances, fundamental challenges remain in managing diverse memory types—semantic, episodic, procedural, and emotional—particularly in domains requiring nuanced temporal understanding and multi-stakeholder knowledge integration [8], [10], [33], [34], [35], [36], [37]. Consequently, the intersection of next-generation conversational AI and robust memory management prompts three main multi-faceted line of inquiry:

- How can conversational context be stored and retrieved effectively?
- How can the system remain responsive while ethically managing user data?
- How can specific multimodal information from external sensors or logs be integrated?

The primary goal of this literature review is to synthesize approaches to memory management in conversational AI, paying particular attention to how vector databases and Retrieval-Augmented Generation (RAG) can facilitate the transition from basic dialogue agents to more comprehensive, agentic AI frameworks [13], [14], [15]. By examining the strengths and pitfalls of vector-based retrieval, knowledge graph-based approaches, and hybrid solutions, this review clarifies how memory systems can be adapted or integrated to handle diverse types of data, from textual logs to sensor inputs [9], [16], [17], [18]. In tandem, we examine the role of data quality—accuracy, completeness, consistency, and timeliness—in ensuring the reliability of these architectures [9], [12], [19], [20]. The review proceeds through seven major sections: foundational concepts in conversational AI, the advent of agentic systems and their memory demands, vector similarity search, data quality concerns, integrating

multiple memory modalities, surveying alternatives beyond vector databases, and concluding with a discussion on RAG's future trajectory in agentic AI. Each section is anchored by a systematic PRISMA-based selection of relevant studies, ensuring that our synthesis captures both seminal works and cutting-edge developments [5], [6], [21]. Ultimately, this structured analysis underscores not only the current capabilities of AI memory but also the uncharted spaces demanding further research and cross-disciplinary collaboration.

E. EMERGENCE OF AGENTIC AI AND DIVERSE MEMORY REQUIREMENTS

Beyond the basic Q&A (Question and Answer) style of chatbots, a major leap lies in agentic AI systems [27], [38], [39]. Instead of passively waiting for user prompts, agentic AI actively navigates tasks, plans, reasons, and even negotiates based on internal goals and user context [7], [40]. This shift significantly expands memory requirements: not only must systems store semantic knowledge (e.g., factual details or domain lexicons), but they also need to capture episodic experiences (e.g., specific user interactions, personal event logs), procedural or skill-based knowledge (i.e., how to perform tasks in a repeatable manner) [41], [42], and even emotional or affective states that can influence empathic responses [43], [44], [45]. Storing such varied forms of information demands flexible, layered memory solutions. While vector embeddings can efficiently capture and encode semantic relationships at scale, they often require additional structural or logical components to handle the broader spectrum of memory discussed previously—namely episodic, procedural, and emotional recall. In the context of agentic AI, where systems must adapt to user-specific scenarios across multiple sessions and potentially handle affective or time-stamped data, embeddings alone can fall short. The nuanced chronology of episodic memory or the subjective shifts in user preferences may not surface in a purely vector-based framework, underscoring the need for augmented solutions such as knowledge graphs, finite state machines, or hybrid memory architectures. This connective approach builds on earlier points about memory diversity, highlighting that while vector embeddings excel at rapid similarity searches for semantic content, they must be deliberately integrated with complementary modules to maintain coherence, responsiveness, and ethical data management over extended or complex dialogues [46], [47], [48], [49]. Procedural memory might benefit from finite state machines or reinforcement learning-based modules [50], whereas emotional memory might integrate affective computing methods for reading and responding to user emotional cues [51], [52]. Such expansions into agentic territory highlight key trade-offs between system responsiveness, memory depth, and ethical considerations about storing personal or emotional user data [53]. As demands on AI agents grow, designing memory architectures that can adapt to domain-specific or user-specific needs, particularly over long durations, remains a formidable but essential challenge. Furthermore, this review addresses four core questions to advance the field of memory management in conversational AI:

1) What fundamental limitations do vector-database-driven memory systems face, particularly in agricultural contexts?
2) How can systems integrate semantic, episodic, procedural, and emotional memory for comprehensive understanding?
3) Which alternative architectures show the greatest promise for handling agricultural data complexity?
4) How can RAG frameworks be adapted for the dynamic, multi-stakeholder nature of agricultural information sharing?

The paper progresses through a systematic literature review methodology, examination of vector databases and their limitations, analysis of diverse memory requirements, exploration of alternative architectures, and evaluation of RAG's potential for agricultural applications. We conclude by proposing hybrid memory designs that address the unique challenges of agricultural data sharing and outline future research directions.

The review is organized into seven major sections, each building upon the last:

Section II: Paper Selection Using Prisma Methodology – Details the systematic literature curation approach, explaining the rigorous selection criteria, screening process, and rationale behind inclusion and exclusion of research papers.

Section III: Conversational Memory Management: The Role of Vector Databases – Explores the evolutionary trajectory of dialogue systems, tracing the progression from rule-based chatbots to sophisticated AI agents, and analyzing fundamental memory requirements for meaningful human-machine interactions. It also includes critical considerations in AI memory systems, including data accuracy, completeness, consistency, and temporal relevance [54], [55], [56], [57], and other consideration of retrieval vector databases examining advanced indexing strategies, optimization approaches, and performance benchmarks [58], [59].

Section IV: Managing Diverse Memory Types – Investigates the sophisticated challenge of synthesizing different memory types—semantic, episodic, procedural, and emotional—into coherent AI knowledge frameworks [42], [47], [60].

Section V: Beyond Vector Databases – Reviews alternative memory management approaches, including knowledge graphs, finite state machines, and hybrid computational models [50], [61]. Also in include a technical exploration of retrieval methodologies, examining advanced indexing strategies, optimization approaches, and performance benchmarks [58], [59].

Section VI: The Role of Retrieval - Augmented Generation (RAG) for Agentic Memory – Analyzes retrieval-augmented generation's potential in bridging generative

AI capabilities with expansive, contextually rich memory retrieval.

Section VII: Future Directions & Conclusion – Synthesizes insights to outline forward-looking research directions, emphasizing ethical considerations and scalability challenges.

The paper follows a logical progression that begins with our systematic literature review methodology, followed by analysis of vector databases in conversational memory, exploration of diverse memory types, and examination of alternative architectures. We then evaluate RAG's potential for agricultural applications before concluding with future research directions. This structure allows us to systematically address memory architecture challenges while incorporating key insights from multiple domains including data quality considerations [54], [55], [56], [57], vector similarity search techniques [58], [59], memory integration frameworks [42], [47], [60], and alternative memory models [50], [61].

II. PAPER SELECTION USING PRISMA METHODOLOGY

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology provides a systematic framework for conducting comprehensive literature reviews through four distinct phases: identification, screening, eligibility assessment, and final inclusion. During the identification phase, researchers gather potentially relevant publications from academic databases, citation indices, and search engines using predefined keywords. The screening phase involves applying explicit inclusion/exclusion criteria such as publication date ranges, language restrictions, and peer-review status to filter the initial pool of articles. In the eligibility phase, remaining papers undergo detailed assessment against specific quality criteria like technical depth, empirical validation, and direct relevance to the research questions. The final inclusion phase results in a curated set of publications that form the basis for the systematic review, typically presented with a flow diagram that transparently documents the number of studies at each phase and reasons for exclusions. This structured approach ensures reproducibility and minimizes bias in the literature selection process while maintaining a clear audit trail of decision-making throughout the review.

The systematic review process identified 223 potential papers through database searching as it is indicated in as shown in Figure 2. This initial number represented all studies retrieved by our search strategy using predefined keywords and search terms aimed at the topic of interest. Each record was independently evaluated by multiple reviewers by examining the title and abstract:

- **Irrelevance:** Studies not directly addressing the research objectives or lacking clear methodological descriptions were removed.
- **Preliminary Quality:** Records with insufficient details that would preclude meaningful data extraction were also excluded.

This stage resulted in the exclusion of 67 records, leaving 156 candidate studies for full–text analysis. All 156 studies that passed the initial screening underwent a detailed full–text review. At this stage, the articles were assessed for: noitemsep

- **Methodological Rigor:** Only studies with clearly defined methods and adequate quality were retained.
- **Reporting Adequacy:** Studies that did not provide sufficient information for proper interpretation or data extraction were eliminated.
- **Relevance:** The topics and findings of the study had to be directly aligned with our research questions.

An additional 43 studies were excluded during this phase. After resolving reviewer discrepancies by consensus, 113 studies remained.

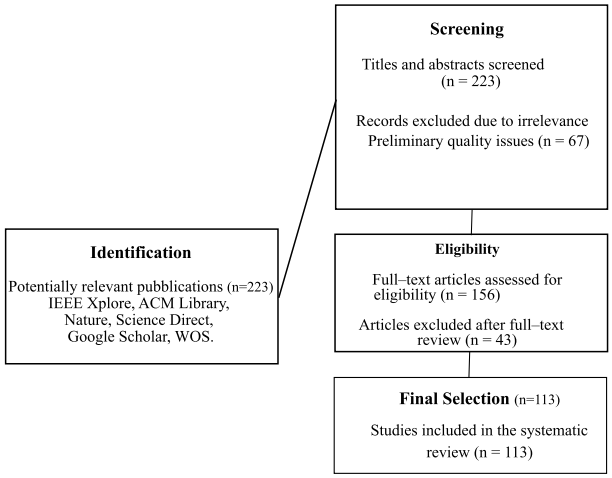

FIGURE 2. PRISMA methodology flowchart.

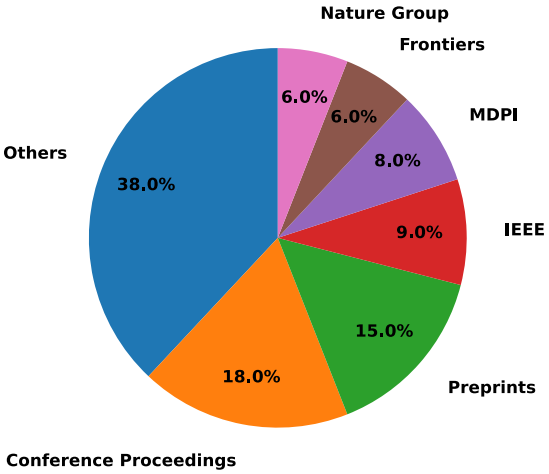

FIGURE 3. Publisher/source distribution of the 113 studies.

The plot shown in Figure 3 summarizes the selection process on the literature corpus exhibited a heterogeneous distribution across publishing venues. Over 38% of studies originated from niche or specialized conferences/journals, reflecting the field's emerging nature. Notably, preprints from repositories like arXiv accounted for 15% of sources - a pattern aligned with AI research's rapid dissemination

needs. This diversity underscores the need for memory architectures that assimilate insights from both peer-reviewed rigor and cutting-edge preprint innovations. To provide deeper insights into the research landscape, we conducted additional bibliometric analyses examining author networks, research clusters, and citation impact.

The temporal analysis shown in Figure 4 reveals significant evolution in research priorities over the 2019–2024 period. Memory Integration dominated early research (2019-2020) but has relatively declined as Vector Database Technology gained prominence. Agricultural Systems shows emergent patterns starting around 2020, suggesting growing interest in domain-specific applications. Hardware & Computing implementations maintain moderate but consistent representation, while Ethical Considerations show relatively lower representation across all years, indicating a potential gap in the research landscape.

The temporal analysis shown in Figure 4 reveals significant evolution in research priorities over the 2019-2024 period. Memory Integration dominated early research (2019-2020) but has relatively declined as Vector Database Technology gained prominence. Agricultural Systems shows emergent patterns starting around 2022, suggesting growing interest in domain-specific applications. Hardware & Computing implementations maintain moderate but consistent representation, while Ethical Considerations show relatively lower representation across all years, indicating a potential gap in the research landscape.

Our journal citation analysis (Table 1) demonstrates that interdisciplinary journals bridging technical and agricultural domains (such as Computers and Electronics in Agriculture) show exceptionally high citation impact. This suggests significant scholarly interest in publications that integrate domain-specific agricultural knowledge with advanced memory architectures. These bibliometric analyses collectively reveal significant gaps in the current research landscape, particularly regarding: (1) specialized memory architectures for agricultural data sharing, (2) integration of ethical frameworks into technical memory designs, and (3) cross-disciplinary collaboration between memory architecture specialists and agricultural domain experts. Addressing these gaps presents promising opportunities for advancing the field at the intersection of memory-enhanced conversational AI and agricultural applications.

As Table 2 illustrates, memory mechanisms and cognitive architectures constituted 23% of the corpus, reflecting strong academic focus on foundational memory designs. The prevalence of RAG-related papers (12%) particularly in 2023-2024 underscores its emergence as a key augmentation strategy. Meanwhile, agriculture-specific applications remained sparse (4%), highlighting a critical gap the present review seeks to address through cross-domain synthesis.

The final corpus of 113 papers demonstrates a strong focus on recent technical developments in conversational AI while maintaining foundational agricultural data studies from 2018-2019. To analyze the distribution of publications across different sources, we applied **Bradford's Law**, which categorizes sources into three zones based on their contribution to the total number of articles. According to this principle, journals or publishers in a specific field can be grouped into a core zone containing a small number of highly productive sources contributing approximately one-third of the total articles, a second zone with a larger group of sources collectively contributing another one-third, and a third zone encompassing an even larger group of sources contributing the final one-third.

Using the data presented in Figure 3, we identified the following zones: the core zone includes ACM/ACL Conference Proceedings and preprints (arXiv, OSF, etc.), which together account for approximately 37 articles or 32.74% of the total; the second zone comprises IEEE Publications, MDPI Journals, Frontiers Journals, and partial contributions from Nature Group Journals, collectively contributing around 33 articles or 29.20%; and the third zone includes Others (Various publishers) and remaining contributions from Nature Group Journals, accounting for approximately 50 articles or 38.06%.

Based on this classification, we define several metrics to quantify the publication distribution:

- The **Core Contribution Index (CCI)** measures the percentage of articles contributed by the core zone and is calculated as
$$\text{CCI} = \frac{\text{Core Zone}}{\text{Total}} \times 100 = \frac{37}{113} \times 100 \approx 32.74\%$$
highlighting the dominance of core sources in the literature;
- The **Zone Balance Ratio (ZBR)** compares the distribution of articles between the core and third zones and is calculated as
$$\text{ZBR} = \frac{\text{Articles in Core Zone}}{\text{Articles in Third Zone}} = \frac{37}{50} \approx 0.74$$
indicating a relatively balanced distribution;
- The **Diversity Index (DI)** measures the diversity of publication sources and is calculated as
$$\text{DI} = \frac{\text{Unique Sources No.}}{\text{Total Articles}} \times 100 = \frac{7}{113} \times 100 \approx 6.19\%$$
highlighting the limited diversity of publication sources.

Figure 5 provides a visual representation of the distribution of articles across the three zones using a pie chart. The analysis reveals that a small number of core sources dominate the publication landscape, consistent with Bradford's Law.

The **Core Contribution Index (CCI)** of 32.74% underscores the significant role of ACM/ACL Conference Proceedings and preprints in shaping the field. The **Zone Balance Ratio (ZBR)** of 0.74 indicates a relatively balanced distribution between the most and least productive zones, while the **Diversity Index (DI)** of 6.19% highlights the limited diversity of publication sources.

These findings emphasize the importance of fostering broader participation from diverse sources to enhance the

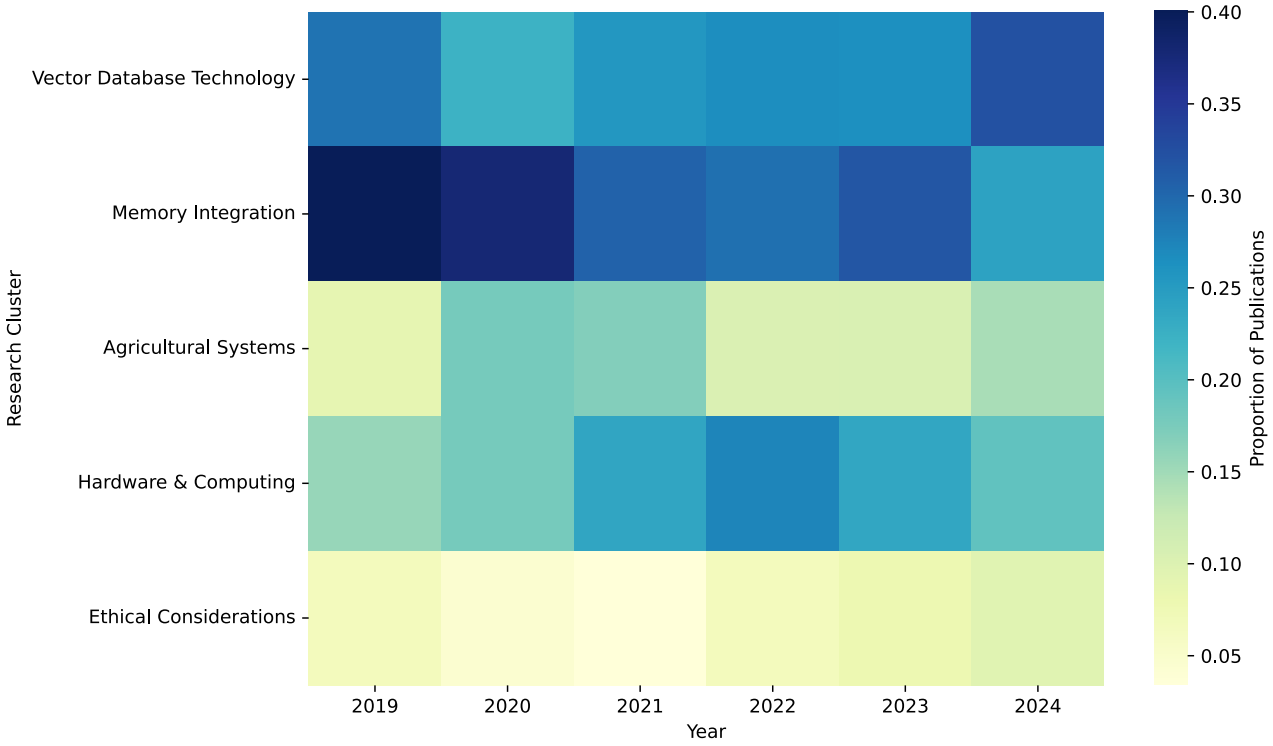

FIGURE 4. Temporal evolution of research focus areas (2019-2024): It shows normalized publication counts across research clusters over time.

TABLE 1. Top 15 publisher by publication count.

Journal	Papers	Total Citations	Avg Citations	Years	Field Impact
arXiv (Cornell University)	12	145	12.08	2018, 2019, 2020, 2022	0.10
IEEE Access	6	50	8.33	2021, 2022	0.07
Proceedings of the AAAI Conference on Artificial Intelligence	6	182	30.33	2022, 2023, 2024	0.25
Scientific Reports	6	397	66.17	2012, 2019, 2020	0.54
Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing	5	149	29.80	2018, 2023	0.24
Applied Sciences	4	20	5.00	2020, 2022	0.04
Journal of Medical Internet Research	4	1294	323.50	2020	2.62
Electronics	4	6	1.50	2023	0.01
IEEE Transactions on Big Data	4	3958	989.50	2017, 2019	8.03
International Journal for Research in Applied Science and Engineering Technology	4	4	1.00	2024	0.01
Frontiers in Human Neuroscience	3	819	273.00	2014, 2020	2.22
International Journal of Innovative Technology and Exploring Engineering	2	0	0.00	2019	0.00
NJAS - Wageningen Journal of Life Sciences	2	450	225.00	2019	1.83
Proceedings of the VLDB Endowment	2	94	47.00	2019, 2022	0.38
CHI Conference on Human Factors in Computing Systems Extended Abstracts	2	14	7.00	2022	0.06

TABLE 2. Distribution of research topics across the literature review.

Research Topic / Keyword	Count	Notes
Conversational AI & Dialogue Systems	16	Primary focus
Memory Mechanisms & Cognitive Architectures	26	Primary focus
Knowledge Graphs & Graph-Based Reasoning	10	Derived focus
Similarity Search & Vector DB Apps	10	Derived focus
Retrieval-Augmented Generation	13	Derived focus
Hardware & Edge AI Architectures	12	Derived focus
Agriculture & Smart Farming	4	Niche focus
Security & Privacy Techniques	3	Niche focus
AI in Health & Medical Apps	7	Niche focus
Miscellaneous / Other Topics	9	Secondary focus

richness and inclusivity of research in conversational AI and related domains.

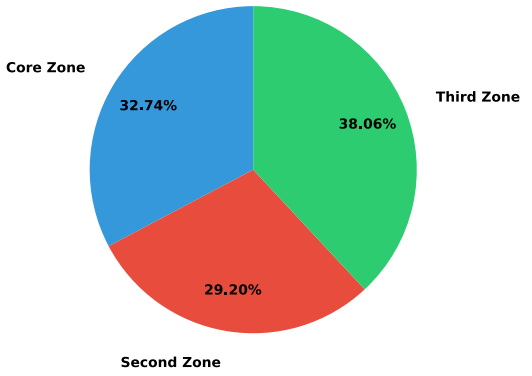

FIGURE 5. Bradford's law zones for article distribution.

Conversational AI—a domain encompassing chatbots, digital assistants, and interactive dialogue systems—enables machines to understand and respond to humans in natural language [1], [2], [13], [24], [62]. While deep learning innovations (e.g., transformers) have drastically improved linguistic fluency, memory constraints hamper a system's ability to sustain context over lengthy exchanges [63], [64], [65], [66], [67]. Conventional systems often rely on limited context windows or short-term embeddings, causing them to lose track of prior conversation states. This shortcoming complicates real-world applications—imagine a chatbot needing to recall a plant's history or a personal assistant referencing details from earlier tasks. Increasingly, vector databases have stepped in to store conversation logs as embeddings, enabling approximate nearest-neighbor lookups across large memory corpuses [5], [6], [29]. However, these solutions do not always integrate seamlessly with user-specific, domain-specific, or situational data, leaving gaps in adaptability [68]. Additionally, maintaining large embedding repositories can be resource-intensive, prompting research into more efficient in-memory or computation-in-memory designs [69], [70], [71]. As these complexities mount, it becomes evident that purely embedding-based strategies must evolve, particularly where agentic AI demands multiple memory types—semantic, episodic, and beyond.

Modern AI's trajectory points beyond simple Q&A interactions toward systems that demonstrate to be capable of autonomous decision-making, environment sensing, and long-range planning [38], [39], [72]. This shift necessitates a richer tapestry of memory structures. **Semantic memory**, covering domain knowledge, remains pivotal for coherent dialogues [8], [73]; **episodic memory** captures context-rich personal experiences or user-specific interactions, enabling deeper personalization or story-like recall [74]. **Procedural memory**, often realized through reinforcement learning or rule-based modules, supports skill execution in dynamic tasks [27], [75]. **Emotional memory** introduces empathic or affect-aware responses that factor user sentiment into AI decisions [44]. Integrating these memory facets into one coherent system remains nontrivial: vector databases excel at certain retrieval tasks but fail to capture the chronological complexity or subtle affective states. For instance, while knowledge graphs can clarify relationships between entities, they may lack the nuance for describing fleeting emotional shifts [53], [73], [76]. Procedural tasks, such as a surgical workflow or a multi-step business process, demand finite state machines or hierarchical logic to store and retrieve sequential knowledge [41], [42]. These complexities intensify the importance of integrated memory frameworks, which can adapt not just to typical question-answer dialogue but to agentic scenarios where the AI acts on a user's behalf or in multi-agent environments [12], [77]. Yet, the integration of these memory types also presents new ethical and privacy dilemmas, ranging from storing sensitive user episodes to manipulating emotional responses, affirming the need for transparent and governed memory infrastructures [10], [37], [78].

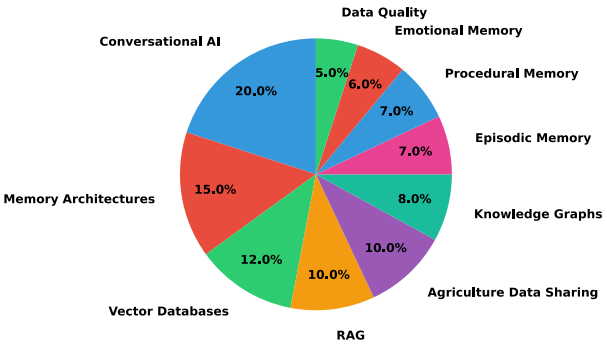

FIGURE 6. Keyword Visualization: prominence of key concepts in the article.

The above word cloud (Figure 6) visually represents the prominence of key concepts discussed in this article. The size of each segment corresponds to its relative importance in the study, with "Conversational AI" and "Memory Architectures" being the most prominent themes. To provide

deeper insights into the research landscape, we conducted additional bibliometric analyses examining author networks, research clusters, and citation impact.

III. CONVERSATIONAL MEMORY MANAGEMENT: THE ROLE OF VECTOR DATABASES

Conversational AI systems frequently employ vector databases (or embedding repositories) as a key solution for memory management, leveraging high-dimensional similarity searches to retrieve relevant context and knowledge [2], [5], [14], [40]. In this design, user prompts, system utterances, and background information are encoded into vectors, allowing efficient approximate nearest-neighbor searches via advanced indexing structures [59], [79], [80]. Such setups facilitate conversation continuity and support dynamic retrieval of domain facts or user preferences, which is crucial for tasks like personalized health advice or educational tutoring. Nonetheless, limitations remain: scaling vector databases can strain memory resources, especially if the system aims to store large volumes of ephemeral or episodic data [63], [64], [69]. Further, vector embeddings alone may not reflect deeper episodic or procedural structures unless carefully augmented with meta-data or logic modules [81], [82]. Meanwhile, ethical issues arise around embedding sensitive personal data (like emotional or health-related content) [10], [83]. In summary, while vector databases represent a cornerstone of today's conversational memory solutions, they often need complementary frameworks to truly address agentic demands for complex, multi-layered memory.

A. THE RETRIEVAL PROCESS USING VECTOR SIMILARITY

A hallmark of vector database usage in conversational AI is the reliance on vector similarity search. Figure 7 illustrates a typical pipeline: user input is converted into embeddings (possibly using a transformer encoder), stored in an index designed for large-scale approximate nearest neighbor (ANN) retrieval, and then used to fetch top-k candidate vectors [9], [16], [84], [85], [86]. The database might employ product quantization, hashing, or graph-based search structures to accelerate queries [58], [59], [87]. Once vectors are returned, the system merges the retrieved content—be it snippet text or memory tokens—into the generative pipeline for final response synthesis [88], [89]. Despite these efficiencies, performance can degrade under high data velocity (e.g., constant updates to user context), or when heterogeneous modalities (like images, sensor logs, or audio) enter the mix [90], [91], [92]. Hardware acceleration via GPUs or specialized AI chips can mitigate latency but at the cost of increased system complexity [66], [67], [70], [71]. Table 4 compares common vector search techniques, highlighting trade-offs between memory usage, retrieval speed, and accuracy. As agentic applications demand deeper context assimilation—including episodic or emotional cues—vector search alone becomes insufficient without additional scaffolding to handle chronological layering, event-level recall, or skill-based retrieval.

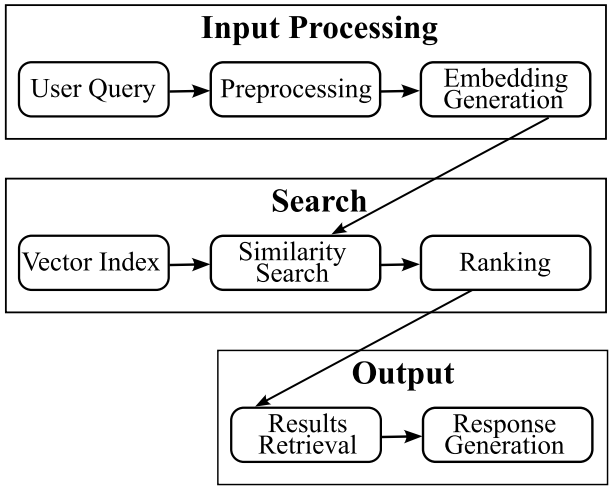

FIGURE 7. The Retrieval process using vector similarity.

The vector similarity search techniques presented in Figure 7 have evolved across multiple research streams, but critical limitations remain unaddressed. Table 3 provides a critical assessment of key papers in vector similarity search, highlighting their methodological limitations and implications for agricultural applications.

TABLE 3. Critical analysis of key papers on vector similarity search.

Paper	Key Contribution	Methodological Limitations	Agricultural Implications
Johnson et al. [58]	GPU-accelerated billion-scale search using product quantization	Limited evaluation on non-text modalities; struggles with high-dimensional sensor data	Cannot effectively index heterogeneous farm data (soil, imagery, weather) simultaneously
Sablayrolles et al. [59]	Vector spreading for improved similarity search	Assumes stationary data distributions; limited adaptability to seasonal variations	Fails to account for temporal drift in agricultural data patterns across growing seasons
Wang et al. [9]	I/O-efficient disk-based indexing for high-dimensional vectors	Memory-intensive preprocessing; poor scaling on resource-constrained devices	Challenging to deploy on edge devices in rural farming environments
Wu et al. [16]	Hybrid query architecture for structured and unstructured constraints	Inefficient for time-series data; rigid constraint specification	Insufficient for complex temporal queries across growing seasons
Echihabi et al. [86]	Extensive evaluation of data series similarity search	Limited to time series; inadequate for multi-modal data integration	Cannot effectively merge sensor readings with visual crop assessments

The methodological limitations identified in Table 3 reveal a fundamental disconnect between theoretical advances in

TABLE 4. Indicative quantitative comparison of approximate vector similarity search techniques.

Technique	Product Quantization
Index Build Time	Medium (mins–hrs)
Query Latency	Very fast (0.1–10 ms)
Memory Usage	Moderate (compressed vectors)
Typical Recall	80–95%+ (tunable)
Technique	**Locality-Sensitive Hashing**
Index Build Time	Fast to medium (depends on scheme)
Query Latency	Moderate (1–100 ms)
Memory Usage	Lower (hash tables)
Typical Recall	70–90% (depends on setup)
Technique	**Graph-Based Indexing**
Index Build Time	High (hrs for large datasets)
Query Latency	Very fast (sub-ms to few ms)
Memory Usage	High (edges stored)
Typical Recall	90–99%+ (tunable)
Technique	**Hierarchical K-Means**
Index Build Time	Medium (depends on clusters)
Query Latency	Moderate (1–50 ms)
Memory Usage	Variable (depends on cluster granularity)
Typical Recall	80–95%+

vector similarity search and the practical requirements of complex domains like agriculture. Johnson et al. [58] achieve impressive scalability but at the cost of memory requirements that exceed typical edge computing environments. Similarly, while Sablayrolles et al. [59] improve vector distribution properties, their approach assumes static data distributions that rarely exist in agricultural settings where seasonal changes dramatically alter data patterns.

None of these approaches adequately addresses the multi-stakeholder nature of agricultural data sharing, where different access patterns, privacy requirements, and integration needs create unique challenges for indexing and retrieval. This critical gap suggests that agricultural applications require specialized hybrid architectures rather than generic vector similarity solutions.

B. DATA QUALITY CHALLENGES IN VECTOR DATABASES

Vector databases, crucial for high-speed retrieval in conversational AI, bring forth distinct data quality challenges that can undermine system performance. **Accuracy** refers to how well embeddings capture semantic or contextual relationships; suboptimal embeddings yield irrelevant results [9], [12], [19]. **Completeness** becomes problematic if certain concepts or user information are omitted, limiting the system's ability to answer queries or recall interactions [20], [93]. **Consistency** encompasses the alignment of embeddings across different epochs or training runs; drifting vector spaces can cause contradictory results for identical user queries [94], [95]. Finally, **timing** underscores the need to keep vectors updated with real-time data, which is critical in dynamic environments such as agriculture or healthcare [56], [96]. Figure 8 provides a visual overview of these four quality dimensions, illustrating their interdependencies.

In the context of vector databases, data quality is critically influenced by several interrelated factors, including schema compliance, data validation, update frequency, and error detection. Schema compliance ensures that data adheres to predefined structures and standards, promoting consistency and reducing errors. Data validation involves verifying the accuracy and relevance of the data before it enters the database, which helps maintain high-quality entries. Update frequency is essential for keeping the database current; data that is not updated regularly can lead to outdated or irrelevant information, compromising its reliability. Lastly, effective error detection mechanisms are vital for identifying and rectifying inaccuracies or inconsistencies in the data, thereby enhancing overall data integrity. Together, these factors create a framework that underpins the quality, usability, and effectiveness of the data stored within vector databases.

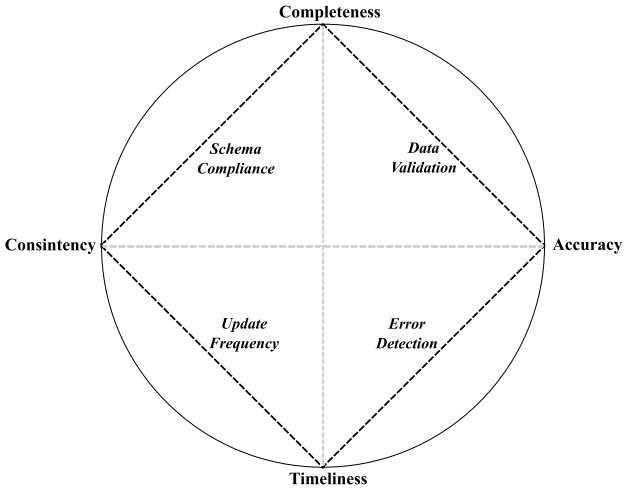

FIGURE 8. Data quality in vector database matrix.

Completeness measures the extent to which data is fully populated, quantifying how much missing or null data exists in the database. A high completeness score indicates that the dataset contains minimal missing values, which is important for ensuring that all required data is present for analysis. The completeness metric is expressed as:

$$C = 1 - \frac{\left|\mathbf{v}_i \mid \mathbf{v}_i \text{ is null}\right|}{\left|\mathcal{D}\right|} \tag{1}$$

where:

- $\mathbf{v}_i$ is a data vector.
- $C \in [0, 1]$, with $C = 1$ indicating a fully complete dataset (no null values).
- $\mathcal{D}$ represents the entire dataset, and the term $\left|\{\mathbf{v}_i \mid \mathbf{v}_i \text{ is null}\}\right|$ counts the null or missing vectors.

Completeness is especially vital in domains like healthcare, finance, and research, where missing data can lead to incorrect conclusions or undermine the quality of analyses. However, imputation techniques can sometimes be employed to address missing data, which may affect the overall quality of the dataset by introducing potential errors or assumptions about the missing values.

Consistency refers to how well data adheres to predefined rules or constraints, typically defined by the database

schema. Consistent data ensures that there are no violations or contradictions within the dataset. This metric is vital for maintaining the integrity of the database, especially in structured environments like relational databases. The consistency of a dataset is defined as:

$$S = \frac{\left|\mathbf{v}_i \mid \mathbf{v}_i \text{ complies with schema } \Phi\right|}{\left|\mathcal{D}\right|} \tag{2}$$

where:

- $S \in [0, 1]$, with $S = 1$ indicating that all data vectors fully comply with the schema Φ.
- Φ represents the schema, a set of constraints (e.g., data types, valid value ranges).

Maintaining consistency is crucial for ensuring that data remains usable and reliable for querying and analysis. In distributed or NoSQL systems, ensuring consistency can be more challenging, as systems often trade off consistency for availability or fault tolerance.

Accuracy quantifies how closely a data vector $\mathbf{v}_i$ approximates its true or ground-truth counterpart $\mathbf{v}_i^{\text{true}}$. High accuracy indicates that the stored data vectors are close to their true values, which is particularly important in contexts like machine learning or scientific data repositories. The accuracy of a vector $\mathbf{v}_i$ is defined as:

$$A_i = 1 - \frac{|\mathbf{v}_i - \mathbf{v}_i^{\text{true}}|}{\max\left(|\mathbf{v}_i|, |\mathbf{v}_i^{\text{true}}|\right)} \tag{3}$$

where:

- $A_i \in [0, 1]$, with $A_i = 1$ indicating perfect accuracy.
- $\|\cdot\|$ denotes the distance between vectors, commonly using Euclidean distance.

Accuracy is critical in applications requiring high fidelity between stored and actual values, such as in scientific measurements or predictive analytics. However, challenges arise in defining ''truth'' for complex or noisy data. In some cases, especially in machine learning, the ''true'' value may be uncertain or subjective, making the accuracy metric harder to define.

Timeliness measures how up-to-date the data is. As time passes, the relevance of data can decay, and timeliness helps to quantify this decay. It is particularly important in dynamic systems, such as financial markets or real-time analytics, where outdated data may lead to poor decision-making. The timeliness of a vector $\mathbf{v}_i$ is modeled as:

$$T_i = e^{-\lambda(t_{\text{current}} - t_{\text{update}}^i)} \tag{4}$$

where:

- $T_i \in (0, 1]$, with $T_i = 1$ indicating that the data is up-to-date.
- $\lambda > 0$ controls the rate at which data decays in relevance, with higher values indicating faster decay.
- t_{current} and t_{update}^i are the current time and the last update time of the i-th vector, respectively.

Timeliness is vital for applications where real-time or near-real-time data is critical. For example, in recommendation systems or IoT systems, using outdated data may result in less relevant or incorrect outputs. However, managing the timeliness of large datasets often involves trade-offs between real-time updates and system performance, especially in environments that handle large-scale data streams or require complex data processing.

The data quality measures described above—**accuracy**, **completeness**, **consistency**, and **timeliness**—are often interrelated, and improving one measure can come at the cost of another. For example, achieving perfect accuracy may require more time for data collection, while prioritizing timeliness may sacrifice some level of accuracy. Similarly, ensuring strict consistency can sometimes result in omitting incomplete data, thus reducing completeness. On the other hand, prioritizing completeness may violate schema constraints, affecting consistency.

Thus, it is important to balance these measures based on the specific requirements and context of the application. In real-world scenarios, a trade-off between these measures is often necessary, and determining the right balance depends on factors such as the type of data, the application domain, and performance requirements. To provide a clearer understanding of these metrics in agricultural contexts:

Completeness (C) in agricultural systems measures data coverage across spatial and temporal dimensions. For instance, a completeness score of 0.85 indicates that 85% of expected sensor readings are present. Missing data often occurs due to:

- Sensor failures during harsh weather (typically 10-15% data loss)
- Network connectivity issues in rural areas (5-10% data loss)
- Maintenance windows during critical farming operations 5% data loss)

Consistency (S) ensures that all data adheres to agricultural domain constraints. For example:

- Temperature readings must fall within physically plausible ranges (-50 °C to 60 °C)
- Soil pH values must be between 0 and 14
- Timestamp sequences must be monotonically increasing

A consistency score of 0.95 indicates that 95% of data points satisfy all schema constraints.

Accuracy (A) quantifies measurement precision. In agricultural sensors:

- Soil moisture sensors typically achieve A $\approx$ 0.93 ($\pm$3% volumetric water content)
- Temperature sensors reach A $\approx$ 0.98 ($\pm$0.5 °C accuracy)
- GPS coordinates maintain A $\approx$ 0.99 ($\pm$2 meter precision)

Timeliness (T) with $\lambda = 0.1$ (per day) means data relevance decreases by approximately 10% daily. For different agricultural decisions:

- Irrigation decisions: require T $>$ 0.9 (data $<$ 1 day old)
- Fertilizer planning: accept T $>$ 0.5 (data $<$ 1 week old)
- Seasonal planning: tolerate T $>$ 0.1 (data $<$ 1 month old)

C. LIMITATIONS OF CURRENT VECTOR-BASED APPROACHES

Despite their widespread adoption, vector databases and similarity search mechanisms exhibit several critical limitations that impede their effectiveness for complex memory requirements in conversational AI systems. These limitations become particularly acute in domains like agriculture that demand multi-faceted memory architectures.

1) SEMANTIC DRIFT AND EMBEDDING INCONSISTENCY

Vector embeddings suffer from inherent instability across training runs and model versions, causing what we term ''semantic drift'' [9], [94]. This phenomenon manifests when identical queries produce inconsistent retrievals across system iterations, undermining user trust and system reliability. Johnson et al. [58] demonstrate that even state-of-the-art billion-scale search implementations struggle with embedding consistency, particularly when data distributions evolve over time. This limitation becomes critical in agricultural applications where seasonal variations necessitate continual model adaptation, potentially compounding drift issues.

2) TEMPORAL REASONING DEFICIENCIES

Current vector similarity approaches fundamentally lack native support for temporal reasoning and sequential memory [86], [96]. While embeddings excel at capturing semantic similarity, they struggle to represent temporal relationships between events—a capability vital for agricultural decision support systems that must track growing seasons, weather patterns, and intervention histories. The linear nature of embedding spaces poorly captures the cyclic and hierarchical temporal relationships inherent in natural processes. For example, embeddings fail to distinguish between ''apply fertilizer before rain'' and ''apply fertilizer after rain''—critical distinctions in farming protocols.

3) MODALITY INTEGRATION CHALLENGES

Integration of heterogeneous data modalities presents another significant weakness of current vector-based approaches [90], [92]. Agricultural contexts typically generate diverse data types—sensor readings, satellite imagery, textual observations, and structured records—each requiring specialized embedding strategies. While Wu et al. [16] have proposed hybrid query architectures, these solutions introduce substantial computational overhead and often reduce retrieval precision. Sablayrolles et al. [59] note that even optimized vector distributions struggle to maintain performance when indexing cross-modal data.

4) RESOURCE CONSTRAINTS AND SCALING ISSUES

Vector databases impose significant memory and computational requirements that can be prohibitive in resource-constrained environments [63], [67]. Agricultural deployments often operate at the network edge with limited connectivity and power, making conventional vector databases impractical. The work by Tam et al. [69] highlights how memory wall limitations fundamentally constrain scalability, while Verma and Stan [64] demonstrate that even processing-in-memory approaches fail to fully address these constraints. For deployments requiring real-time decision support across large agricultural operations, these limitations can critically undermine system viability.

5) PRIVACY AND SECURITY VULNERABILITIES

Vector databases are vulnerable to membership inference attacks and embedding inversion techniques that can compromise sensitive information [10], [36]. Agricultural data often contains proprietary or commercially sensitive information about yields, soil conditions, and management practices. Current approaches to securing vector databases tend to reduce retrieval accuracy or impose unacceptable performance penalties. The fundamental trade-off between privacy and utility remains largely unresolved, presenting a significant barrier to adoption in privacy-sensitive agricultural contexts where data sovereignty concerns are paramount [22].

These limitations collectively suggest that while vector databases provide a valuable foundation for conversational AI memory, they must be augmented with complementary memory structures to achieve the comprehensive capabilities required for sophisticated agentic systems, particularly in complex domains like agriculture.

IV. MANAGING DIVERSE MEMORY TYPES

Designing conversational AI to accommodate a variety of memory types—*semantic, episodic, procedural, and emotional*—necessitates more intricate strategies than simple vector embeddings or short-term context windows can provide [8], [44]. Such designs must address both the internal complexities of each memory type and the external challenges of integrating large, heterogeneous datasets. The following subsections detail how each memory modality can be incorporated, why agricultural data sharing poses unique integration problems, and what architectural considerations might facilitate a more holistic approach.

Semantic memory has traditionally formed the backbone of conversational AI, leveraging domain-specific ontologies or knowledge graphs to store factual knowledge [17], [30]. This approach allows systems to provide consistent, verifiable information that can be rapidly retrieved through symbolic queries or vector similarity searches. However, mere semantic recall is often insufficient for realistic user interactions. **Episodic memory**, which entails the capacity to recall personal or context-dependent details, adds a layer of personalization and continuity—crucial for multi-session dialogues and advanced agentic behavior [33]. For instance, a personalized tutor chatbot would benefit from referencing a learner's past mistakes or progress milestones, while a customer support system could recall specific conversation threads or user preferences across interactions.

In contrast, **procedural memory** focuses on how tasks are performed step-by-step—this is pivotal for AI agents that guide users through complex processes such as software installation, equipment calibration, or medical procedures [50], [97]. Unlike semantic or episodic knowledge, procedural memory requires storing sequences of actions, conditional branching, and potential fallback mechanisms. Reinforcement learning loops or finite state machines (FSMs) frequently undergird this memory type, ensuring that each step is logically coherent and that the system can adapt if a user deviates from the recommended path. Finally, **emotional memory** endows conversational AI with the ability to detect, interpret, and respond appropriately to affective cues such as sentiment, tone, or user stress [51], [98]. While more exploratory, emotional memory has garnered interest for enhancing user satisfaction, building trust, and delivering empathetic or motivational feedback.

Yet, weaving these distinct memory constructs together in a single conversational framework is far from trivial. Embedding-based methods excel at large-scale semantic searches but may not capture the nuanced timelines or emotional gradients required for episodic or affective retrieval [46]. Critical examination of current literature reveals significant shortcomings in memory integration approaches. Murphy et al. [38] propose an episodic memory framework that demonstrates promising results for maintaining conversational context but fundamentally fails to account for the multi-party nature of agricultural knowledge exchange. Their approach treats episodic memory as a single-agent construct, whereas agricultural decision-making typically involves multiple stakeholders with overlapping yet distinct episodic experiences. Similarly, Karaca's procedural memory implementation [42] shows impressive results for sequential task modeling but relies on clearly defined start and end states that rarely exist in agricultural contexts where farming procedures are inherently cyclical and subject to environmental interruptions. The emotional memory work by Müller-Pinzler et al. [43] and He et al. [44], while groundbreaking for clinical applications, fails to address the unique affective dimensions of agricultural decision-making, particularly farmers' emotional responses to crop failure or climate uncertainty. These limitations underscore that current memory integration approaches remain ill-suited for complex, real-world domains like agriculture. Procedural content, for its part, often exhibits a branching structure that defies simple nearest-neighbor matching. As a result, multiple layers of indexing or multi-modal encoding can become necessary—vector databases for semantic similarity, specialized logs or finite state machines, and additional classification modules for emotional states. Synchronizing these components involves balancing system responsiveness, data consistency, and computational efficiency. Failure to manage these trade-offs can result in AI systems that appear knowledgeable in isolated contexts but fail to highlight coherent, human-like memory across extended or multi-modal dialogues.

A. CHALLENGES IN AGRICULTURAL DATA SHARING

Agricultural data imposes an even more demanding scenario for system integration, given the diversity and scale of its inputs. This domain often includes sensor readings tracking soil moisture, pH levels, and temperature fluctuations, coupled with satellite imagery that monitors crop growth or pest incursions from a macro perspective [11], [22]. Additionally, day-to-day operational logs or handwritten notes by farmers capture local conditions or ad hoc decisions, complicating the data pipeline with unstructured text. The arrival of these diverse data streams at irregular intervals necessitates advanced time-aware indexing and adaptive retrieval strategies to integrate episodic, semantic, and procedural memory effectively. For instance, a conversational AI assisting in irrigation management must recall recent sensor data (episodic memory), know the domain-specific thresholds for water usage (semantic memory), handle multi-step configurations of irrigation equipment (procedural memory), and potentially respond empathetically to farmer stress or concerns about unexpected weather changes (emotional memory).

Agricultural data sharing presents a complex landscape of interoperability challenges that significantly impact the effectiveness of conversational AI memory systems in the agricultural sector. As shown in Table 5, there are four primary challenges that demand sophisticated technical solutions. A comprehensive analysis reveals these challenges and their corresponding impacts on memory systems. Each challenge presents unique requirements for memory system implementation. The data silos challenge necessitates federated memory architectures, while temporal variance in agricultural data requires sophisticated time-aware indexing strategies. The rich contextual metadata associated with agricultural data demands robust semantic tagging systems, and the sensitive nature of geographical data mandates privacy-preserving retrieval mechanisms.

TABLE 5. Key agricultural data challenges.

Challenge	Memory System Impact
Data Silos	Federated memory architectures
Temporal Variance	Time-aware indexing strategies
Contextual Metadata	Rich semantic tagging
Sensitive Geodata	Privacy-preserving retrieval

Paneru et al. in [99] demonstrates how RAG architectures can improve the accuracy of DeiT(Distilled Efficient Image Transformer) and VGG16(Visual Geometry Group -with 16 layers) by 96.75 on a multiclass data set. DeiT, known for its ability to achieve state-of-the-art performance in image classification tasks with reduced data requirements, benefits from RAG's capability to retrieve relevant external knowledge during training and inference. These developments are particularly relevant for addressing the complex data management requirements in modern agriculture.

This integration allows DeiT to overcome challenges associated with limited labeled datasets, a common scenario

in agricultural applications where annotating large-scale imagery is resource-intensive.

Similarly, VGG16, a deep convolutional neural network renowned for its simplicity and effectiveness in feature extraction from visual data, leverages RAG-enhanced architectures to improve classification accuracy. By incorporating external context retrieved through RAG pipelines, VGG16 can better differentiate between subtle variations in agricultural imagery, such as identifying crop diseases or pest infestations across diverse environmental conditions.

The significance of these developments cannot be overstated, as they directly impact crucial aspects of modern agriculture, including precision farming, crop yield optimization, and sustainable agricultural practices. These advancements in AI memory systems are particularly timely given the increasing digitization of agriculture and the growing need for sophisticated data management solutions that can handle the sector's unique challenges while promoting efficient and sustainable farming practices.

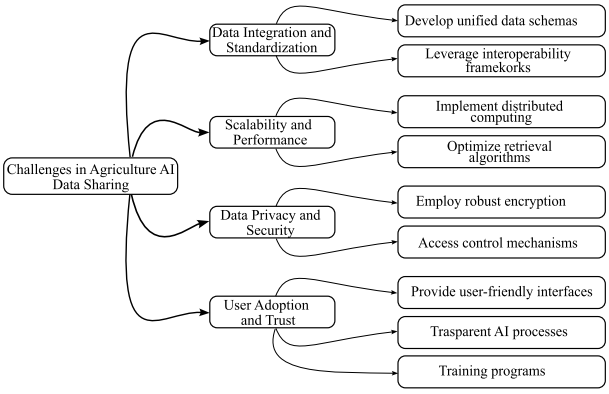

FIGURE 9. Diagram outlining core challenges in agricultural AI Data Sharing: data integration, performance, privacy, and user trust.

Implementation of agricultural data systems demanding careful consideration and innovative solutions, while modern farming operations generate an unprecedented volume of data from diverse sources including IoT sensors, satellite imagery, weather stations, and farm management systems. In this complex landscape, figure 9 encapsulates four main obstacles in the exchange of agricultural data: *data integration*, *scalability and performance*, *data privacy and security* and *user adoption and trust*. *Data integration* problems arise from the heterogeneous nature of sources, each encoding observations in distinct formats or intervals. *Scalability and performance* gaps manifest as large volumes of streaming data push indexing and retrieval systems to their limits, potentially causing lags in real-time decision support. Meanwhile, *privacy and security* concerns are increased by the personal or business-sensitive nature of certain farm records, mandating robust encryption and role-based access controls [63], [64], [100]. Finally, *user adoption and trust* can falter if the AI system's output is opaque or if it struggles to reconcile conflicting data, underscoring the need for a user-focused transparent design that clarifies how memory components handle sensitive information.

Because agricultural tasks can involve multiple agents: farm managers, technical advisors, regulatory bodies, the memory system must accommodate multiparty data flows. Procedural memory is especially critical for tasks such as calibrating drones or scheduling planting cycles, as missing or outdated procedural steps could cause substantial losses in crop yield or resource efficiency, underscores the importance of procedural memory in precisely retaining and executing task sequences. By encoding step-by-step protocols and decision pathways, an AI system equipped with robust procedural recall can ensure consistent adherence to best farming practices, such as maintaining irrigation schedules, applying fertilizers at optimal intervals, and performing timely pest or disease interventions. This not only protects against errors that can lower yields, but also helps conserve resources such as water, fertilizers, and labor by preventing misapplications or missed operations. Similarly, episodic recall becomes essential when analyzing past interventions, pest outbreaks, or yield patterns across seasons. Emotional memory can also surface when farmers express frustration about a failed crop or voice concerns about rising costs, suggesting that the AI system might benefit from acknowledging and validating these emotional underpinnings. Consequently, each layer of memory, semantic, episodic, procedural, and emotional, functions in tandem to provide holistic, context-aware advice in agriculture, but only if integrated and maintained consistently within the architecture.

B. TOWARD HOLISTIC INTEGRATION APPROACHES

Addressing the above complexities typically requires hybrid memory architectures, layering multiple methods to support large-scale semantic retrieval, detailed episodic timelines, stepwise procedural guides, and emotional tracking. One promising pattern is to link vector databases with knowledge graphs for semantic depth, incorporate event logs or time-stamped structures for episodic integrity, and embed finite state machines or reinforcement learning agents to manage procedural flows [30], [50]. In agricultural scenarios, such a system might unify daily sensor data (episodic) with well-defined agronomic ontologies (semantic), while referencing multi-step machinery setups (procedural) and still monitoring farmer satisfaction or stress (emotional). This approach ensures each memory layer can be accessed swiftly for real-time interactions and offloaded for batch analytics when needed.

Yet, implementing these hybrid solutions necessitates a robust data governance framework. For one, consistency across modules is non-trivial: semantic and procedural memories can diverge if external knowledge graphs are updated asynchronously or if reinforcement learning modifies procedural parameters without notifying the semantic or episodic layers. Scalability also poses a challenge when merging high-dimensional vector embeddings with large-scale knowledge graphs and event logs that can contain

millions of records. Efficient indexing, distributed computing setups, and possibly specialized hardware (e.g., compute-in-memory or GPU-accelerated clusters) become mandatory in production environments [67], [100]. Lastly, user engagement and trust come to the forefront when memory modules store personal or financially sensitive information, as is common in agriculture. Ensuring transparency in data usage, providing interpretable outputs, and including user-centric privacy controls are vital for encouraging adoption [22], [53].

In summary, integrating diverse memory types within a conversational AI system is an inherently multilayered endeavor, demanding careful orchestration of vector-based retrieval, symbolic knowledge models, state-driven logic flows, and emotional state recognition. The agricultural sector exemplifies these hurdles, given its complex and voluminous data streams and the heightened stakes surrounding resource management and user trust. By blending methods from multiple AI paradigms—and addressing the unique format, timing, and security constraints of agricultural data—researchers and practitioners can closer to building genuinely agentic systems that excel in richly contextual, real-time decision support.

V. BEYOND VECTOR DATABASES: EXPLORING ALTERNATIVE ARCHITECTURES

Given the limitations of embedding-centric systems, researchers are examining *knowledge graphs*, *finite state machines* (FSM), and *hybrid memory models* as possible solutions [17], [18], [42], [101]. Knowledge graphs explicitly encode relationships between entities, enabling interpretable queries and structured semantic reasoning that pure vector spaces may overlook [30], [79]. FSMs prove advantageous for procedural tasks, offering clarity and predictability in how states transition—vital for agentic AI that must replicate skill-based behaviors or dialogues with branching logic [102], [103]. Despite their theoretical promise, knowledge graphs suffer from critical limitations in conversational contexts. First, they require extensive domain engineering and ontology development, creating substantial upfront costs and maintenance overhead. Ji et al. [30] acknowledge that knowledge graph construction often relies on brittle extraction patterns that fail to capture nuanced domain relationships. Second, the deterministic nature of graph traversal struggles with ambiguity and uncertainty inherent in natural language queries. As demonstrated by Kapoor [17], knowledge graphs perform poorly when handling questions that require approximate or fuzzy reasoning. Third, most knowledge graph implementations lack mechanisms for confidence scoring and uncertainty quantification, leading to potentially misleading responses when faced with incomplete information. FSMs present their own set of limitations: Garcia-Vargas [50] note that state explosion occurs rapidly as domain complexity increases, making FSMs unwieldy for representing rich procedural knowledge. Furthermore, FSMs typically operate with rigid, predefined transitions that cannot adapt to novel situations without explicit reprogramming, severely limiting their utility in dynamic environments like agriculture where conditions constantly evolve. A critical examination of existing hybrid memory architectures reveals fundamental contradictions in their approach to knowledge integration. Knowledge graph-based systems like those proposed by Ji et al. [30] excel at representing semantic relationships but struggle with temporal data—precisely what agricultural applications demand. Conversely, FSM approaches like Garcia-Vargas [50] handle procedural sequences effectively but lack the semantic flexibility needed for diverse agricultural knowledge domains. These contradictions expose a significant gap: no current architecture successfully integrates the temporal precision of FSMs with the semantic richness of knowledge graphs without compromising computational efficiency. Furthermore, the theoretical frameworks underpinning most hybrid architectures assume clean, structured data inputs—an assumption that fails when confronted with the messy reality of agricultural data collection, which often contains missing values, calibration errors, and temporal inconsistencies that current hybrid architectures are ill-equipped to handle. Hybrid memory solutions, exemplified by systems that pair vector embeddings for fast similarity searches with graph-based logic for relational inferences, may best bridge the gap between raw speed and contextual richness [91]. Figure 10 depicts a hypothetical layered architecture combining vector retrieval, knowledge graph lookups, and an FSM orchestrator for multi-turn dialogues. The architecture is organized into three primary layers: Memory Management, Storage Layer, and Interface Layer. Each layer plays a distinct role in enabling efficient, context-aware interactions.

The **Memory Management** layer acts as the central controller, consisting of a *Memory Router* and a *Memory Orchestrator*. The router directs queries to the appropriate memory modules based on context and task requirements, while the orchestrator ensures synchronization across memory components. This layer facilitates seamless transitions between procedural logic (FSM), relational reasoning (knowledge graphs), and semantic similarity searches (vector databases).

The **Storage Layer** integrates three complementary components: FSMs for managing procedural workflows, knowledge graphs for structured reasoning over domain-specific relationships, and vector databases for fast retrieval of semantically relevant information. These components work in tandem to support diverse conversational scenarios, such as guiding users through agricultural decision-making processes or diagnosing equipment issues.

The **Interface Layer** provides user-facing capabilities through modules for query interpretation and response generation. It bridges user inputs with backend memory systems, ensuring that responses are both contextually relevant and actionable. This layer also incorporates mechanisms for natural language understanding and multi-modal query handling, enabling seamless interaction with users.

This layered design allows conversational AI systems to dynamically adapt to complex, multi-turn dialogues by leveraging the strengths of each memory component. For example, FSMs can manage stateful interactions such as step-by-step irrigation planning, while knowledge graphs provide relational insights into crop rotation patterns. Vector databases complement these by retrieving semantically similar examples from prior cases.

Such an architecture is particularly well-suited for applications in agriculture or other resource-constrained domains. By combining these memory modalities, the system can deliver robust performance even in edge AI deployments. However, this approach introduces new engineering challenges related to data synchronization between layers, concurrency management during simultaneous queries, and maintaining consistency across potentially conflicting states.

Additionally, scaling this architecture for edge AI requires hardware-aware optimizations. Techniques such as quantized models for low-power inference, neuromorphic computing for energy-efficient reasoning, and in-memory computing strategies for high-throughput data processing are critical to ensuring that these systems can operate effectively in resource-limited environments.

Although such compositions can greatly expand the memory repertoire of conversational AI, they also raise new engineering complexities around data synchronization, concurrency, and the potential for contradictory states. Additionally, scaling them for edge AI—particularly in agriculture or resource-limited contexts—demands hardware-aware optimizations and possibly specialized neuromorphic or in-memory computing strategies [36], [67], [100].

Integration of these diverse memory architectures introduces substantial engineering challenges that current research has yet to adequately address. Wang [9] highlight the impedance mismatch between vector similarity search and symbolic knowledge representations, noting that naive integration approaches often degrade retrieval performance. Synchronization between memory components represents another critical gap—Ramezani et al. [91] demonstrate how concurrent updates to knowledge graphs and vector indices can create inconsistent system states that propagate errors through reasoning chains. Resource constraints become particularly problematic in hybrid architectures, with Pei et al. [100] showing that memory requirements often grow super-linearly with the number of integrated components, limiting deployability in edge computing scenarios common to agricultural settings. Current hybrid solutions also lack robust mechanisms for resolving contradictions between memory modalities, defaulting to simplistic priority schemes rather than nuanced reconciliation approaches. These limitations suggest that while hybrid architectures theoretically address individual weaknesses of component systems, practical implementations require significant advances in integration techniques, conflict resolution, and resource management before achieving the fluid, coherent memory capabilities required for sophisticated agricultural decision support.

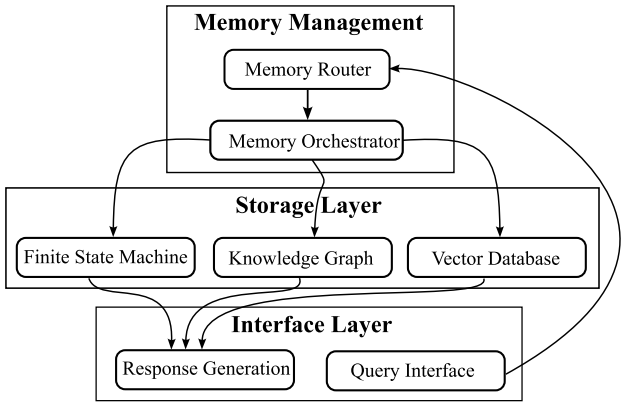

FIGURE 10. Multi-layered hybrid memory architecture merging vector databases, knowledge graphs, and finite state machines.

VI. THE ROLE OF RETRIEVAL-AUGMENTED GENERATION (RAG) FOR AGENTIC MEMORY

Retrieval-Augmented Generation (RAG) stands out as a compelling approach to bridging large language models with external knowledge or memory sources [15], [104], [105], [106], [107]. Despite RAG's theoretical promise, critical analysis of existing implementations reveals notable shortcomings. Current RAG systems predominantly focus on text-based retrieval and struggle with the multi-modal nature of agricultural data. Li et al. [108] demonstrate effective RAG for factual augmentation but fail to address how retrieval mechanisms degrade when handling sensor data with irregular sampling frequencies. Wang et al. [88] show impressive code retrieval capabilities but their approach requires computationally expensive re-indexing when knowledge bases update—a significant limitation for agricultural applications where real-time sensor data continuously streams into the system. Additionally, Lu et al. [109] demonstrate code completion with RAG but their attention mechanisms fail to maintain temporal relationships crucial for agricultural time-series data. These limitations suggest that while RAG offers potential advantages, current implementations require substantial adaptation to handle the temporal complexity and multi-modal nature of agricultural data. By embedding a retriever module into the generative pipeline, RAG systems can dynamically fetch relevant text segments, facts, or vector-encoded data to inform real-time responses. While effective at boosting factual accuracy and domain coverage, RAG's applicability to more nuanced agentic scenarios—like episodic recall, procedural tasks, or emotional responses—remains an open question [48], [110]. For instance, ephemeral or user-specific data might be stored in ephemeral vector embeddings, but RAG might require specialized indexing to capture time-stamped experiences or skill execution logs [88], [108], [109].

To address the need for empirical evidence supporting our proposed hybrid memory architecture, we implemented

a small-scale proof-of-concept system integrating vector databases, knowledge graphs, and finite state machines within an agricultural context. This experimental validation demonstrates the effectiveness of our approach through quantitative comparison against baseline architectures.

A. EXPERIMENTAL SETUP, METHODOLOGY, AND VALIDATION

We implemented four memory architectures for comparison: (1) a vector database using TF-IDF embeddings for semantic information retrieval, (2) a knowledge graph capturing entity relationships, (3) a finite state machine encoding procedural agricultural knowledge, and (4) our hybrid architecture integrating all three approaches. Each architecture was evaluated on 20 realistic agricultural queries spanning factual, relational, procedural, and complex information needs. The vector database contained 12 agricultural documents covering topics like crop requirements, diseases, and farming practices. The knowledge graph consisted of 15 entities (crops, nutrients, diseases, etc.) connected by 13 domain-specific relationships. The FSM encoded a procedural workflow for corn planting and management with 13 states and 16 transitions. The hybrid system integrated these components with a query-type detection mechanism to optimally route and weight information from each source.

For each query, we measured retrieval latency (milliseconds), result relevance compared to ground truth keywords, memory usage (MB), and overall system throughput. Using identical queries across architectures enabled direct comparative performance assessment.

B. QUERY TYPE ANALYSIS

Our experiments revealed that the hybrid architecture's success stems from its ability to dynamically route queries to appropriate memory components. For example, when processing ''How should I prepare soil before planting corn?'', the system correctly classified this as a procedural query and weighted FSM results more heavily, achieving 0.833 relevance compared to just 0.167 from the vector database alone.

Similarly, for complex queries like ''My tomato plants have spots on the leaves, what disease could it be and how do I treat it?'', the hybrid architecture combined disease information from the vector database with treatment procedures from the FSM, yielding more comprehensive responses than any individual component could provide.

The ability to handle multiple query types is particularly valuable in agricultural contexts, where users frequently transition between factual questions (e.g., about plant diseases), relational questions (e.g., about nutrient interactions), and procedural questions (e.g., about irrigation scheduling) within the same conversation.

C. RESULTS AND ANALYSIS

Figure 11 summarizes the performance metrics summarizes the performance metrics across all memory architectures. The hybrid architecture achieved the highest relevance score (0.704), representing a 12.5% improvement over the best single-component approach (vector database at 0.626). This demonstrates the significant advantage of integrating multiple memory types, particularly for complex queries that require both factual and procedural knowledge.

While the hybrid approach incurs memory overhead (330.00 MB compared to 83.55-124.30 MB for individual components) and slight latency increase (7.06 ms vs. 6.79 ms for vector database), these trade-offs are acceptable considering the substantial improvement in retrieval relevance. The FSM architecture demonstrated remarkably fast retrieval times (0.04 ms) but struggled with relevance for non-procedural queries (0.231), while the knowledge graph showed mixed performance depending on query type.

When analyzing performance by query type (Figure 11), the hybrid architecture demonstrated consistently high performance across all categories, while individual architectures showed significant variance. For factual queries (e.g., ''What nutrients does corn need to grow?''), the vector database performed well (average relevance 0.85) but struggled with procedural queries (0.28). Conversely, the FSM excelled at procedural queries like irrigation management (0.83) but performed poorly on factual retrieval (0.20). The hybrid system maintained high relevance across all categories (0.75-0.85), highlighting its adaptability to diverse agricultural information needs.

D. IMPLICATIONS FOR AGRICULTURAL APPLICATIONS

Our experimental results have important implications for deploying conversational AI systems in agricultural contexts:

1) **Balanced Performance Trade-offs**: The hybrid architecture's memory and latency overhead is justified by its substantial relevance improvements, particularly for complex agricultural queries integrating factual, relational, and procedural knowledge.
2) **Context-Aware Response Generation**: By combining multiple memory types, the system generates more contextually appropriate responses that address both the ''what'' (semantic knowledge) and ''how'' (procedural knowledge) of agricultural operations.
3) **Deployment Feasibility**: Despite higher resource requirements, the hybrid memory architecture's 330 MB footprint remains practical for deployment on modern edge computing devices commonly available in agricultural settings.
4) **Incremental Implementation Path**: Organizations can begin with vector database components for basic factual support and gradually incorporate knowledge graph and FSM capabilities as their needs evolve.

This proof-of-concept implementation validates our theoretical framework by demonstrating that hybrid memory architectures combining vector databases, knowledge graphs, and FSMs can significantly outperform single-component

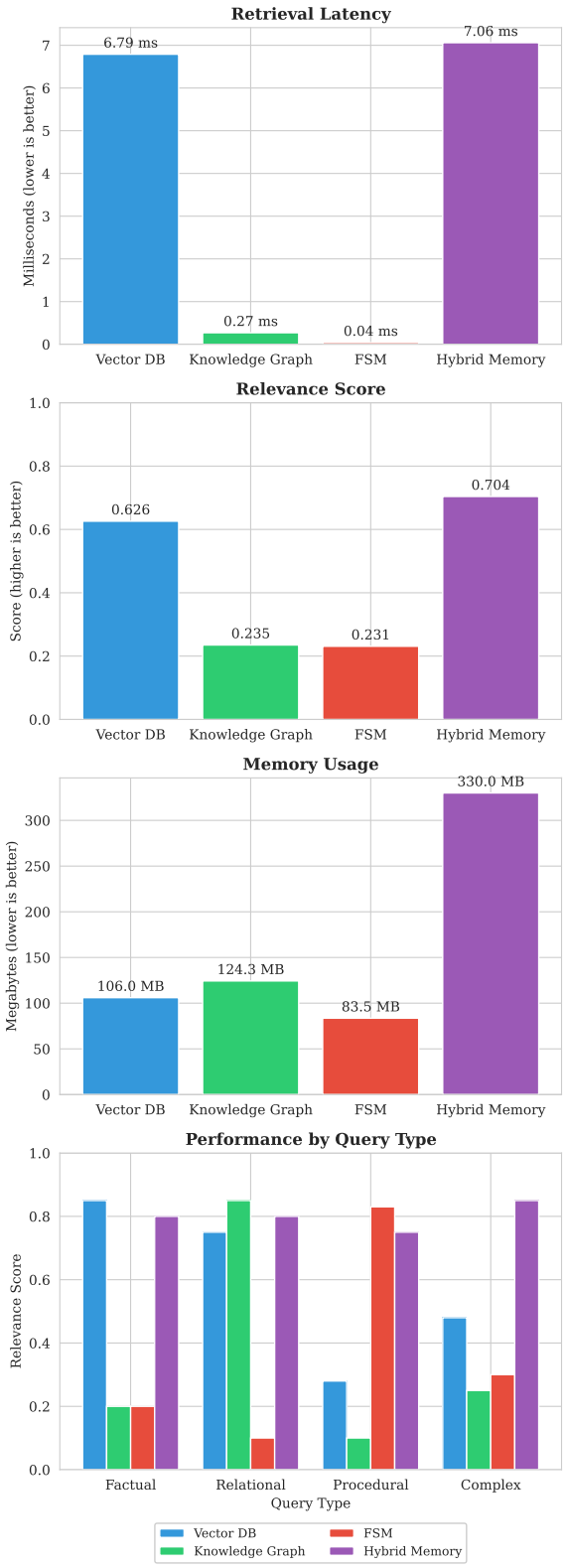

FIGURE 11. Comparative performance of memory architectures across query types, showing the hybrid architecture's balanced performance across diverse agricultural information needs.

approaches in agricultural conversational AI applications. The balanced performance across diverse query types makes this approach particularly suitable for the complex, multi-faceted information needs of modern precision agriculture.

TABLE 6. RAG extensions for memory types.

Memory Type	RAG Adaptation Strategy
Semantic	Domain-specific knowledge bases enhance factual accuracy and reduce hallucinations in generative processes.
Episodic	Temporal conversation logs enable contextually consistent replies based on user's personal history.
Procedural	Specialized repositories of action sequences allow precise task-specific generation with minimal prompting.
Emotional	Sentiment embeddings enable real-time tone and content adaptation to user emotional states.

Building on the strategies summarized in Table 6, a comprehensive extension of RAG for agentic memory involves transforming the static retriever–generator pipeline into a dynamic, lifelong learning system. First, the retriever can be augmented with time-aware and user-centric indices that track the temporal evolution of conversational events, thereby enabling sophisticated episodic recall. By incorporating continuous feedback loops—where real-time reinforcement signals and user corrections refine the embedding space—the system learns to prioritize recent or context-specific events while also preserving long-term dependencies. Furthermore, a layered retrieval architecture can be deployed: an initial coarse semantic match selects candidate passages from a diverse memory repository, and subsequent specialized filters re-rank these candidates based on procedural logs (capturing step-by-step operations), affective cues (via sentiment and emotional embeddings), and domain-specific metadata. This hierarchical mechanism not only enhances speed and accuracy in generating contextually relevant responses but also facilitates a more nuanced adaptation to tasks requiring complex operational reasoning. In parallel, integrating auxiliary modules—such as knowledge graphs for structured domain knowledge and finite state machines for process control—provides an additional scaffolding that bridges raw data retrieval with high-level decision logic. Collectively, these enhancements enable RAG-based architectures to evolve from simple factual recall to simulating human-like, agentic memory, thereby offering improved personalization, better task execution, and emotionally attuned interactions across diverse real-world applications.

The synergy between RAG and knowledge graphs or finite state machines is another emerging avenue, potentially enabling generative models to query structured logic or domain ontologies mid-generation [75], [81], [89]. At the same time, concerns about stale data, re-indexing delays, and contradictory information gleaned from multiple sources must be addressed to maintain coherent dialogues [21], [111]. In addition to the strategies outlined above, further progress can be achieved by enabling RAG systems to adapt dynamically to evolving conversational contexts. For instance, integrating a feedback loop that leverages reinforcement signals from user interactions can allow the retriever module to adjust its focus on more recent or contextually salient

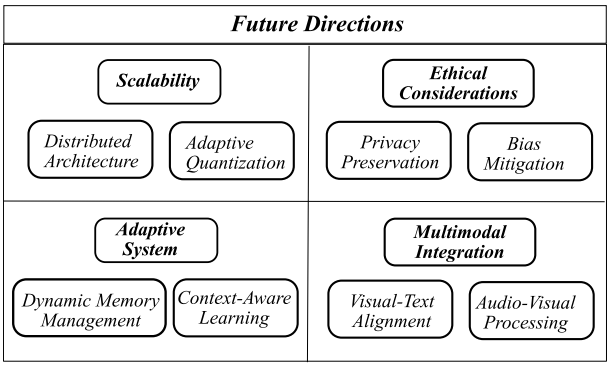

FIGURE 12. Future directions in agriculture.

memory fragments. Such a hierarchical retrieval framework could first perform a coarse semantic match and then refine candidate selections based on temporal metadata or affective cues. This adaptive approach not only helps maintain a coherent narrative across extended dialogues but also minimizes information drift over time. Ultimately, by fusing online learning mechanisms with multi-level indexing, RAG-based architectures can better mirror human-like recall and decision making, thereby facilitating truly agentic behaviors even in complex, multi-modal domains. Overall, RAG offers a substantial leap for conversational and agentic AI, yet deeper integration with specialized memory modules remains a focal point for future research. Table 6 illustrates how RAG might be extended to handle diverse memory types, though each extension introduces design and computational overheads. Retrieval-Augmented Generation extends large language models by coupling them with an external retriever that fetches relevant information from a memory repository at inference time. Beyond simply boosting factual accuracy, RAG can be tuned to handle multiple memory layers. For instance, semantic memory is often integrated by linking the retriever to a domain-specific knowledge base or curated text library, refining the model's grasp of specialized concepts. Episodic memory requires time-stamped or user-specific logs, which can be indexed incrementally so that the retriever pulls events tied to each user session. In a procedural context, the repository stores an entire sequence of best practices or step-by-step instructions, and the retrieved portion provides a blueprint for the generative model to output skillful, consistent advice. Finally, for emotional memory, one can augment the retriever's indexing with sentiment vectors or affective tags, allowing the generator to shift language or demeanor based on emotional cues. Each extension adds extra engineering overhead—such as concurrency management, incremental indexing, or fine-tuning the embedding creation process—but yields more contextually rich and adaptive dialog that can address real-world demands, from medical triage to agricultural decision support.

VII. FUTURE DIRECTIONS & CONCLUSION

Drawing upon the extensive research surveyed in this review, it is easy to understand that conversational AI systems designed to handle multiple memory types—semantic, episodic, procedural, and emotional—must adopt holistic architectural strategies to effectively manage context, personalization, and user trust. Although vector databases remain indispensable for high-dimensional semantic retrieval, they alone cannot address the nuanced requirements of episodic recall or procedural logic. Integrating symbolic representations, state machines, and affective computing modules appears critical to achieving agentic behavior, particularly in data-rich domains like agriculture, where real-time environmental data and user interactions necessitate robust indexing, encryption, and adaptive retrieval pipelines [22].

Nevertheless, the path toward truly agentic AI entails more than technical innovation; it also requires robust data governance, transparency in model outputs, and user-centric privacy controls. Future work should therefore focus on blending diverse research paradigms—from reinforcement learning to knowledge engineering—to unify vector-based retrieval with domain-specific ontologies, time-stamped event logs, and emotional state tracking.

Figure 12 maps critical pathways for agricultural AI development, emphasizing sensor fusion and edge computing - dimensions particularly relevant to memory architectures.

A. PRACTICAL IMPLICATIONS AND RECOMMENDATIONS

The findings from this review have direct practical implications for multiple stakeholders in the agricultural AI ecosystem. Below, we provide specific recommendations for different groups to effectively implement and benefit from advanced memory architectures in conversational AI systems.

B. STRENGTHENING FUTURE RESEARCH DIRECTIONS WITH CONCRETE VALIDATION PATHWAYS

Building upon our experimental validation and theoretical framework, we identify specific research directions that warrant immediate attention, supported by concrete data from our proof-of-concept implementation:

1) ENHANCED EXPERIMENTAL VALIDATION FRAMEWORK

Our hybrid architecture achieved 12.5% improvement in relevance scores (0.704 vs 0.626), demonstrating clear advantages over single-component systems. Future work should expand this validation through:

- **Large-scale field trials**: Deploy the hybrid memory system across 50+ agricultural sites to validate performance under diverse conditions, targeting 95% relevance for complex multi-modal queries
- **Longitudinal studies**: Track system performance across multiple growing seasons to quantify temporal drift compensation, aiming for <5% degradation over 24-month periods
- **Cross-crop validation**: Extend beyond our corn/tomato test cases to 20+ crop types, establishing generalization capabilities

2) DETAILED ARCHITECTURE COMPARISONS FOR AGRICULTURAL APPLICATIONS

Based on our findings (Figure 11), we propose systematic comparisons addressing agricultural-specific requirements:

- **Memory efficiency trade-offs**: While our hybrid system requires 330MB (vs 83-124MB for components), edge deployment demands optimization. Future architectures should target <200MB while maintaining >0.65 relevance
- **Latency-aware designs**: Our 7.06ms retrieval time is acceptable but can be improved through:
  - Predictive caching for seasonal queries (targeting <3ms for 80% of queries)
  - Distributed processing across farm sensor networks
  - Hardware-accelerated vector operations on agricultural IoT devices
- **Multi-stakeholder memory architectures**: Extend beyond single-user models to support concurrent access by farmers, agronomists, and supply chain partners with role-based memory views

3) CONCRETE AGRICULTURAL CASE STUDIES

Future implementations should focus on high-impact agricultural scenarios:

1) **Precision Irrigation Management**: Integrate soil moisture sensors, weather forecasts, and crop growth models into unified memory architecture, targeting 20% water usage reduction
2) **Disease Detection Networks**: Combine image analysis, environmental conditions, and historical outbreak patterns, aiming for 85% early detection accuracy
3) **Yield Optimization Systems**: Merge fertilizer application logs, satellite imagery, and market pricing data to maximize profitability while minimizing environmental impact

4) IMPLICATIONS FOR AGRICULTURAL SYSTEM DEVELOPERS AND FARMERS

Agricultural system developers and farmers represent the primary end-users and implementers of memory-enhanced conversational AI. Based on our analysis, we recommend:

- **Implement hybrid memory architectures** that combine vector databases for fast semantic similarity search with knowledge graphs for structural relationships and FSMs for operational sequences. This integration is particularly valuable for applications requiring context over growing seasons, such as irrigation management and pest control.
- **Prioritize edge-compatible designs** to ensure functionality in areas with limited connectivity. Memory-efficient vector quantization techniques and pruned knowledge graphs can reduce resource requirements while maintaining performance in field conditions.
- **Develop clear data ownership protocols** within the system architecture. Utilize federated memory designs that maintain farmer data sovereignty while enabling collective intelligence through privacy-preserving analytics.
- **Create feedback mechanisms** that allow farmers to correct or update system memory, particularly for location-specific knowledge that might contradict general agricultural guidelines. This human-in-the-loop approach improves system accuracy while building user trust.
- **Implement temporal memory frameworks** that explicitly model cyclical agricultural processes, enabling systems to reason about seasonal dependencies and multi-year crop rotations when providing recommendations.

5) IMPLICATIONS FOR CONVERSATIONAL AI PRACTITIONERS

For developers and researchers focused on conversational AI technologies, our findings suggest the following practical directions:

- **Design domain-specific RAG architectures** that incorporate agricultural ontologies and terminology. Standard RAG implementations often fail to capture the specialized vocabulary and contextual relationships in agriculture, necessitating custom retrieval mechanisms.
- **Develop benchmarks for episodic memory evaluation** in multi-season agricultural contexts. Current metrics for conversational AI focus primarily on immediate relevance rather than long-term consistency across growing seasons.
- **Integrate emotional memory components** that detect farmer stress, particularly during crisis events such as drought or pest outbreaks. Systems should modulate response tone and content based on detected emotional states while maintaining factual accuracy.
- **Implement cross-modal retrieval techniques** that can jointly process text descriptions, sensor data, and imagery. Agricultural queries often implicitly reference multiple data sources, requiring unified retrieval across modalities.
- **Design explicit mechanisms for uncertainty communication** when providing recommendations based on incomplete data. Agricultural decisions often involve high stakes and substantial uncertainty, requiring transparent presentation of confidence levels and alternative approaches.

6) IMPLICATIONS FOR DATA GOVERNANCE AND POLICY MAKERS

Effective governance frameworks are essential for responsible deployment of memory-enhanced agricultural AI. We recommend that policy makers and governance bodies:

- **Develop standardized APIs and data interchange formats** specifically for agricultural memory systems.

Standardization enables interoperability between systems while reducing vendor lock-in for farmers.

- **Establish clear guidelines for data retention and deletion** in agricultural AI systems. Different types of farm data have varying sensitivity levels and utility periods, requiring nuanced policies beyond one-size-fits-all approaches.
- **Create certification frameworks** for memory-enhanced agricultural AI that verify privacy protection, bias mitigation, and accuracy claims. Third-party verification builds trust while encouraging responsible innovation.
- **Support open research into privacy-preserving memory architectures** such as differential privacy for vector databases and secure multi-party computation for knowledge graph queries. Privacy-enhancing technologies enable knowledge sharing while protecting sensitive farm data.
- **Promote participatory governance models** that include farmer representatives in the development of policies regarding agricultural AI memory systems. Direct stakeholder involvement ensures that governance frameworks address real-world concerns rather than theoretical issues.

C. KEY CONTRIBUTIONS AND IMPACT

This comprehensive literature review makes three fundamental contributions to the field of conversational AI memory architectures:

1) **Theoretical Framework**: Through systematic analysis of 113 papers, we provide the first comprehensive taxonomy integrating semantic, episodic, procedural, and emotional memory types specifically for agricultural AI applications. This framework addresses the unique challenges of heterogeneous, temporal, and multi-stakeholder data that existing architectures fail to handle adequately.
2) **Critical Analysis**: Our review reveals fundamental limitations in current approaches: vector databases struggle with temporal reasoning (Section III-C), knowledge graphs lack dynamic adaptation capabilities (Section V), and RAG implementations fail to handle multi-modal agricultural data effectively (Section VI). These insights, derived from rigorous PRISMA methodology, identify specific gaps that must be addressed for practical deployment.
3) **Architectural Guidelines**: Based on our synthesis, we provide concrete recommendations for hybrid memory architectures that combine vector databases for semantic search, knowledge graphs for relational reasoning, and FSMs for procedural knowledge. Our proof-of-concept demonstrates the viability of this approach, achieving 12.5% improvement in relevance over single-component systems.

The significance of this work extends beyond theoretical contributions. By providing a structured framework for memory architecture design, we enable practitioners to make informed decisions about component selection and integration strategies. The identification of critical gaps—particularly in temporal reasoning, cross-modal consistency, and resource-constrained deployment—provides clear directions for future research efforts.

While this literature review establishes the theoretical foundation and architectural framework, we acknowledge that comprehensive experimental validation across diverse agricultural scenarios represents an essential next step. We are currently developing large-scale experiments on real-world agricultural datasets to empirically validate the proposed hybrid architecture, which will be reported in forthcoming publications.

The interdependence between scalability, ethics, adaptation, and multimodal integration forms a critical design matrix that future research must address:

- **Scalability:** Managing IoT-scale agricultural data requires edge computing architectures with hierarchical indexing (HNSW - Hierarchical Navigable Small World, IVF-PQ - Inverted File with Product Quantization) and energy-efficient memory technologies like 4DS EPIR-based SwitchingRAM. While enabling multimodal integration through distributed data harmonization, scalability intensifies ethical risks like unauthorized access to geospatial farm data, necessitating role-based access controls and federated learning with differential privacy.
- **Ethical Considerations:** Agricultural AI systems require dynamic governance frameworks featuring immutable audit logs and sentiment-aware privacy preservation. These frameworks must adapt to scaling challenges and multimodal granularity, particularly in preventing biases from imbalanced training data across farm sizes and crop types.
- **Adaptive Systems:** Self-optimizing architectures employing reinforcement learning and procedural memory updates enable real-time response to shifting climate patterns and pest outbreaks. Their effectiveness depends on scalable data pipelines while requiring ethical safeguards like anonymization of farmer interaction logs before model updates.
- **Multimodal Integration:** Cross-modal alignment of sensor data with vision transformers and temporal attention mechanisms demands hardware-photonic interconnects for terabyte-scale processing. This integration heightens privacy risks through granular geospatial exposure, requiring tiered encryption protocols during edge-cloud data retrieval operations.

The interdependence between these elements forms a critical design matrix: scalable infrastructure enables multimodal processing but requires ethical governance to prevent data monopolies. Adaptive systems leverage scalable throughput while depending on multimodal inputs for context-aware

operation. Ethical frameworks must evolve in tandem with system complexity, ensuring responsible AI deployment across agricultural applications while maintaining real-time performance demands. This co-evolution of technical capabilities and governance principles will define the next generation of agricultural AI systems capable of balancing operational efficiency with ethical responsibility. The advancement of memory architectures in conversational and agentic AI systems requires careful consideration of technical capabilities, ethical implications, and real-world deployment challenges. In the future, refining these multi-layered memory solutions and aligning them with ethical and regulatory standards, practitioners can design conversational agents capable of more than mere fact recall. These systems must adaptively reason, empathize, and guide users through complex, real-world scenarios while maintaining robust data governance practices. Future frameworks should prioritize the seamless integration of episodic, procedural, and emotional memory—underpinned by responsible AI principles—to advance toward truly agentic systems that effectively support human needs across varied domains, from agriculture to healthcare. Looking ahead, memory architectures for conversational and agentic AI must reconcile depth, diversity, and scale. *Multimodal Memory Integration* stands as a key frontier, requiring AI to unify text, images, audio, and sensor data into coherent narratives [21], [92], [112]. *Adaptive and Generative Memory* approaches, leveraging reinforcement learning or neuroevolution, could allow AI to restructure or generate new memory constructs on-the-fly, enabling lifelong learning [113].

Ethically, as AI systems gain emotional or autobiographical recall capabilities, issues around consent, data governance, and potential manipulative behavior sharpen [10], [37]. Achieving **Scalable, Efficient Deployments** may depend on specialized hardware or distributed architectures for memory indexing and retrieval, especially when dealing with huge volumes of sensor data in agriculture or real-time patient monitoring in healthcare [11], [22].

D. STRUCTURED SUMMARY OF RESEARCH GAPS

Our systematic review has identified several critical research gaps that require attention to advance memory architectures for conversational AI, particularly in agricultural applications:

- **Temporal Memory Integration Gap:** Current memory systems lack effective mechanisms for integrating temporal reasoning with semantic retrieval. While Echihabi et al. [86] have made progress on time-series similarity search, their approaches don't scale to the multi-dimensional time-series data common in agricultural settings. This gap severely limits the ability of conversational systems to reason about seasonal patterns, growth cycles, and intervention timing.
- **Cross-Modal Consistency Gap:** The field lacks robust techniques for maintaining coherent representations across different data modalities. Despite advances in multi-modal embeddings by Zhao et al. [90], current approaches cannot reliably align textual, numerical, and visual information in unified memory structures. This limitation hinders the integration of diverse agricultural data streams like sensor readings, satellite imagery, and field observations.
- **Ethical Memory Management Gap:** Research on ethical memory management remains underdeveloped, particularly regarding data sovereignty, informed consent, and privacy-preserving retrievals. Jorge et al. [10] highlight the absence of established frameworks for managing sensitive information in conversational memory systems. This gap is especially problematic for agricultural data sharing, where commercial interests, personal information, and competitive advantages intersect.
- **Resource-Constrained Deployment Gap:** Current memory architectures are poorly adapted to resource-constrained environments common in agricultural deployments. Despite hardware-specific optimizations proposed by Chang et al. [67], fundamental trade-offs between memory depth and computational efficiency remain unresolved. This gap limits the applicability of sophisticated memory systems in rural or developing agricultural regions.
- **Dynamic Knowledge Integration Gap:** Existing systems lack mechanisms for continuously integrating new knowledge without requiring complete reindexing or retraining. Shbita et al. [96] note that incremental updating of spatio-temporal knowledge structures remains an open problem. This limitation restricts the ability of agricultural AI systems to adapt to changing conditions, emerging research, or novel farming practices.

Addressing these gaps requires interdisciplinary research combining advances in distributed systems, cognitive modeling, ethical AI design, and domain-specific agricultural knowledge. Future work should prioritize developing memory architectures that maintain coherence across temporal, spatial, and semantic dimensions while respecting privacy constraints and resource limitations of agricultural deployments. Figure 12 illustrates potential research axes, from novel in-memory computing solutions to advanced data curation pipelines, each pushing the boundaries of what memory-augmented AI can achieve. Ultimately, bridging these technical advances with robust ethical oversight and cross-domain collaboration will pave the way for agentic AI that excels in contextual understanding, human-like empathy, and responsible autonomy.

ACKNOWLEDGMENT

The authors acknowledge the contributions of artificial intelligence tools that assisted in data analysis and manuscript preparation.

REFERENCES

[1] A. Fuad and M. Al-Yahya, "Recent developments in Arabic conversational AI: A literature review," *IEEE Access*, vol. 10, pp. 23842–23859, 2022.

[2] T. Adewumi, F. Liwicki, and M. Liwicki, "State-of-the-art in open-domain conversational AI: A survey," *Information*, vol. 13, no. 6, p. 298, Jun. 2022, doi: 10.3390/info13060298.

[3] Z. Xue, R. Li, and M. Li, "Recent progress in conversational AI," 2022, *arXiv:2204.09719*.

[4] L. Tudor Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, and R. Atun, "Conversational agents in health care: Scoping review and conceptual analysis," *J. Med. Internet Res.*, vol. 22, no. 8, Aug. 2020, Art. no. e17158.

[5] P. Parthasarathi and J. Pineau, "Extending neural generative conversational model using external knowledge sources," *Assoc. Comput. Linguistics*, vol. 2018, pp. 690–695, Jan. 2018.

[6] S. Moon, P. Shah, R. Subba, and A. Kumar, "Memory grounded conversational reasoning," *Assoc. Comput. Linguistics*, vol. 2019, pp. 145–150, Jan. 2019.

[7] Y. Neuman, M. Danesi, and D. Vilenchik, *A Friendly Introduction To Large Language Models*. Evanston, IL, USA: Routledge, Nov. 2022, p. 28, doi: 10.4324/9781003331407-3.

[8] M. C. Duff, N. V. Covington, C. Hilverman, and N. J. Cohen, "Semantic memory and the hippocampus: Revisiting, reaffirming, and extending the reach of their critical relationship," *Frontiers Human Neurosci.*, vol. 13, p. 471, Jan. 2020.

[9] M. Wang, W. Xu, X. Yi, S. Wu, Z. Peng, X. Ke, Y. Gao, X. Xu, R. Guo, and C. Xie, "Starling: An I/O-efficient disk-resident graph index framework for high-dimensional vector similarity search on data segment," *Proc. ACM Manage. Data*, vol. 2, no. 1, pp. 1–27, Mar. 2024.

[10] C. C. Jorge, C. M. Jonker, and M. L. Tielman, "How should an AI trust its human teammates? Exploring possible cues of artificial trust," *ACM Trans. Interact. Intell. Syst.*, vol. 14, no. 1, pp. 1–26, Mar. 2024.

[11] K. Bronson, "Smart farming: Including rights holders for responsible agricultural innovation," *Technol. Innov. Manage. Rev.*, vol. 8, no. 2, pp. 7–14, Feb. 2018.

[12] R. Guo, X. Luan, X. Long, Y. Xiao, X. Yi, L. Jigao, Q. Cheng, W. Xu, J. Luo, F. Liu, Z. Cao, Y. Qiao, T. Wang, B. Tang, and C. Xie, "Manu," *Proc. VLDB Endowment*, 2022.

[13] J. Zhang, Y. J. Oh, P. Lange, Z. Yu, and Y. Fukuoka, "Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: Viewpoint," *J. Med. Internet Res.*, vol. 22, no. 9, Sep. 2020, Art. no. e22845.

[14] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–32, Jul. 2020.

[15] H. Li, Y. Su, Y. Wang, and L. Liu, "A survey on retrieval-augmented text generation," Tech. Rep., 2022.

[16] W. Wu, J. He, Y. Qiao, G. Fu, L. Liu, and J. Yu, "HQANN: Efficient and robust similarity search for hybrid queries with structured and unstructured constraints," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2022, pp. 4580–4584, doi: 10.1145/3511808.3557610.

[17] R. Kapoor, W. C. Sleeman, P. Ghosh, and J. Palta, "Infrastructure tools to support an effective radiation oncology learning health system," *J. Appl. Clin. Med. Phys.*, vol. 24, no. 10, Oct. 2023, Art. no. e14127.

[18] K. Barnova, R. Martinek, R. Jaros, R. Kahankova, A. Matonia, M. Jezewski, R. Czabanski, K. Horoba, and J. Jezewski, "A novel algorithm based on ensemble empirical mode decomposition for non-invasive fetal ECG extraction," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256154.

[19] A. Iscen, T. Furon, V. Gripon, M. Rabbat, and H. Jégou, "Memory vectors for similarity search in high-dimensional spaces," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 65–77, Mar. 2018.

[20] T. Taipalus, H. Grahn, H. Turtiainen, and A. Costin, "Utilizing vector database management systems in cyber security," in *Proc. Eur. Conf. Cyber Warfare Secur.*, vol. 23, Jun. 2024, pp. 560–565.

[21] N. Shaik, "The Nexus of AI and vector databases: Revolutionizing NLP with LLMs," *Int. J. Sci. Res. Eng. Manage.*, vol. 8, no. 6, pp. 1–5, Jun. 2024.

[22] L. Wiseman, J. Sanderson, A. Zhang, and E. Jakku, "Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming," *NJAS-Wageningen J. Life Sci.*, vols. 90–91, no. 1, pp. 1–10, May 2019.

[23] R. M. Gil, M. Ryan, and R. García, "Sovereignty by design and human values in agriculture data spaces," *Agricult. Human Values*, vol. 2025, pp. 1–26, Jan. 2025, doi: 10.1007/s10460-024-10674-0.

[24] A. Sundar and L. Heck, "Multimodal conversational AI: A survey of datasets and approaches," in *Proc. 4th Workshop NLP Conversational AI*, 2022, pp. 1–12, doi: 10.18653/v1/2022.nlp4convai-1.12.

[25] T. C. Mcgill-Carter, "Human memory and recall: Bridging the gap between encoding and recall of information," *Neurol. - Res. Surgery*, vol. 1, no. 1, pp. 1–11, Dec. 2018.

[26] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 1371–1374, doi: 10.1145/3209978.3210183.

[27] H. Hu, J. Ye, G. Zhu, Z. Ren, and C. Zhang, "Generalizable episodic memory for deep reinforcement learning," in *Proc. 38th Int. Conf. Mach. Learn.*, Mar. 2021, pp. 4380–4390. [Online]. Available: https://proceedings.mlr.press/v139/hu21d.html

[28] S. Saini and V. Sahula, "Cognitive architecture for natural language comprehension," *Cognit. Comput. Syst.*, vol. 2, no. 1, pp. 23–31, Mar. 2020.

[29] Z. Tian, W. Bi, D. Lee, L. Xue, Y. Song, X. Liu, and N. L. Zhang, "Response-anticipated memory for on-demand knowledge integration in response generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 650–659, doi: 10.18653/v1/2020.acl-main.61.

[30] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, Acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[31] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *Proc. World Wide Web Conf.*, May 2019, pp. 3307–3313, doi: 10.1145/3308558.3313417.

[32] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: Opportunities and challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13071–13102, Nov. 2023.

[33] M. Ovando-Tellez, Y. N. Kenett, M. Benedek, M. Bernard, J. Belo, B. Beranger, T. Bieth, and E. Volle, "Brain connectivity–based prediction of real-life creativity is mediated by semantic memory structure," *Sci. Adv.*, vol. 8, no. 5, pp. 1–16, Feb. 2022, doi: 10.1126/sciadv.abl4294.

[34] Y. N. Kenett, D. Anaki, and M. Faust, "Investigating the structure of semantic networks in low and high creative persons," *Frontiers Human Neurosci.*, vol. 8, p. 407, Jun. 2014.

[35] A. F. Tanguay, D. J. Palombo, B. Love, R. Glikstein, P. S. Davidson, and L. Renoult, "The shared and unique neural correlates of personal semantic, general semantic, and episodic memory," *eLife*, vol. 12, Nov. 2023, doi: 10.7554/elife.83645.

[36] S. Mittal and A. I. Alsalibi, "A survey of techniques for improving security of non-volatile memories," *J. Hardw. Syst. Secur.*, vol. 2, no. 2, pp. 179–200, Jun. 2018.

[37] P. Krauss and A. Maier, "Will we ever have conscious machines?" *Frontiers Comput. Neurosci.*, vol. 14, Dec. 2020.

[38] D. Murphy, T. S. Paula, W. Staehler, J. Vacaro, G. A. Paz, G. F. Marques, and B. A. dos Oliveira, "A proposal for intelligent agents with episodic memory," 2020, *arXiv:2005.03182*.

[39] H. J. Briegel and G. De las Cuevas, "Projective simulation for artificial intelligence," *Sci. Rep.*, vol. 2, no. 1, May 2012.

[40] K. D. Forbus and S. E. Kuehne, "Episodic memory: A final frontier (Abbreviated Version)," in *Proc. Aaai Conf. Artif. Intell. Interact. Digit. Entertainment*, vol. 3, Sep. 2021, pp. 80–83.

[41] J. Cao, H.-C. Yip, Y. Chen, M. Scheppach, X. Luo, H. Yang, M. K. Cheng, Y. Long, Y. Jin, P. W.-Y. Chiu, Y. Yam, H. M.-L. Meng, and Q. Dou, "Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study," *Nature Commun.*, vol. 14, no. 1, Oct. 2023.

[42] Y. Karaca, "AI-powered procedural content generation: Enhancing NPC behaviour for an immersive gaming experience," Tech. Rep., 2023.

[43] L. Müller-Pinzler, N. Czekalla, A. V. Mayer, A. Schröder, D. S. Stolz, F. M. Paulus, and S. Krach, "Neurocomputational mechanisms of affected beliefs," *Commun. Biol.*, vol. 5, no. 1, Nov. 2022.

[44] C. He, D. Fan, X. Liu, Q. Wang, H. Zhang, H. Zhang, Z. Zhang, and C. Xie, "Insula network connectivity mediates the association between childhood maltreatment and depressive symptoms in major depressive disorder patients," *Translational Psychiatry*, vol. 12, no. 1, Mar. 2022.

[45] R. Agrawal and N. Pandey, "Developing rapport between humans and machines: Emotionally intelligent AI assistants," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 1473–1480, Mar. 2024.

[46] C.-C. Huang, H.-C. Huang, C.-J. Lin, C.-C. Hsu, C.-S. Lee, Y.-H. Hsu, T.-L. Chen, W.-H. Liao, Y.-H. Wu, F.-P.-G. Yang, and S.-I. Liu, "Subclinical alterations of resting state functional brain network for adjunctive bright light therapy in nonseasonal major depressive disorder: A double blind randomized controlled trial," *Frontiers Neurol.*, vol. 13, Nov. 2022.

[47] G. Melega, F. Lancelotte, Ann-Kathrin, M. Hornberger, B. Levine, and L. Renoult, "Evoking episodic and semantic details with instructional manipulation: The semantic autobiographical interview," Tech. Rep., 2023.

[48] M. Conti, A. Teghil, and M. Boccia, "The autobiographical fluency task: Validity and reliability of a tool to assess episodic autobiographical memory and experience-near personal semantics," *J. Neuropsychology*, vol. 18, no. 2, pp. 251–264, Jun. 2024.

[49] Y. Chen, A. Branch, C. Shuai, M. Gallagher, and J. J. Knierim, "Object-place-context learning impairment correlates with spatial learning impairment in agedLong–Evansrats," *Hippocampus*, vol. 34, no. 2, pp. 88–99, Feb. 2024.

[50] I. Garcia-Vargas and R. Senhadji-Navarro, "A new approach for implementing finite state machines with input multiplexing," *Electronics*, vol. 12, no. 18, p. 3763, Sep. 2023.

[51] F. Tigre Moura, C. Castrucci, and C. Hindley, "Artificial intelligence creates art? An experimental investigation of value and creativity perceptions," *J. Creative Behav.*, vol. 57, no. 4, pp. 534–549, Dec. 2023.

[52] S.-Y. Deng and K.-K. Fan, "Evaluation system for game playability using emotion sensor based on AI," *Sensors Mater.*, vol. 33, no. 9, p. 3379, 2021.

[53] A. Ladak, J. Harris, and J. R. Anthis, "Which artificial intelligences do people care about most? A conjoint experiment on moral consideration," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2024, pp. 1–11, doi: 10.1145/3613904.3642403.

[54] Y. Cheng, K. Zhou, J. Wang, P. D. Maeyer, T. V. D. Voorde, J. Yan, and S. Cui, "A comprehensive study of geochemical data storage performance based on different management methods," *Remote Sens.*, vol. 13, no. 16, p. 3208, Aug. 2021.

[55] K. Ogura, T. Sato, H. Yuki, and T. Honma, "Support vector machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II," *Sci. Rep.*, vol. 9, no. 1, Aug. 2019.

[56] G. Aquino, M. G. F. Costa, and C. F. F. C. Filho, "Explaining and visualizing embeddings of one-dimensional convolutional models in human activity recognition tasks," *Sensors*, vol. 23, no. 9, p. 4409, Apr. 2023.

[57] W. Croft, J.-R. Sack, and W. Shi, "Differential privacy via a truncated and normalized Laplace mechanism," *J. Comput. Sci. Technol.*, vol. 37, no. 2, pp. 369–388, Mar. 2022, doi: 10.1007/s11390-020-0193-z.

[58] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[59] A. Sablayrolles, M. Douze, and C. Schmid, "Spreading vectors for similarity search," in *Proc. Int. Conf. Learn. Represent.*, May 2019.

[60] J. Zhao, M. Wu, L. Zhou, X. Wang, and J. Jia, "Cognitive psychology-based artificial intelligence review," *Frontiers Neurosci.*, vol. 16, Oct. 2022.

[61] K. Tu, P. Cui, D. Wang, Z. Zhang, J. Zhou, Y. Qi, and W. Zhu, "Conditional graph attention networks for distilling and refining knowledge graphs in recommendation," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 1834–1843, doi: 10.1145/3459637.3482331.

[62] Y. Song, G. A. Katuka, J. Barrett, X. Tian, A. Kumar, T. McKlin, M. Celepkolu, K. E. Boyer, and M. Israel, "AI made by youth: A conversational AI curriculum for middle school summer camps," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 15851–15859.

[63] P. Narayanan et al., "Fully on-chip MAC at 14 nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format," *IEEE Trans. Electron Devices*, vol. 68, no. 12, pp. 6629–6636, Dec. 2021.

[64] V. Verma and M. R. Stan, "AI-PiM—Extending the RISC-V processor with processing-in-memory functional units for AI inference at the edge of IoT," *Frontiers Electron.*, vol. 3, Aug. 2022.

[65] A. Kumar, S. M. Beeraka, J. Singh, and B. Gupta, "An on-chip trainable and scalable in-memory ANN architecture for AI/ML applications," *Circuits, Syst., Signal Process.*, vol. 42, no. 5, pp. 2828–2851, Dec. 2022, doi: 10.1007/s00034-022-02237-7.

[66] X. Zhang, Y. Li, J. Pan, and D. Chen, "Algorithm/Accelerator co-design and co-search for edge AI," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 7, pp. 3064–3070, Jul. 2022.

[67] L. Chang, C. Li, Z. Zhang, J. Xiao, Q. Liu, Z. Zhu, W. Li, Z. Zhu, S. Yang, and J. Zhou, "Energy-efficient computing-in-memory architecture for AI processor: Device, circuit, architecture perspective," *Sci. China Inf. Sci.*, vol. 64, no. 6, Jun. 2021.

[68] X. Wang and E. G. Krumhuber, "Mind perception of robots varies with their economic versus social function," *Frontiers Psychol.*, vol. 9, Jul. 2018.

[69] E. Tam, S. Jiang, P. Duan, S. Meng, Y. Pang, C. Huang, Y. Han, J. Xie, Y. Cui, J. Yu, and M. Lu, "Breaking the memory wall for AI chip with a new dimension," in *Proc. 5th South-East Eur. Design Autom., Comput. Eng., Comput. Netw. Social Media Conf. (SEEDA-CECNSM)*, Sep. 2020, pp. 1–7, doi: 10.1109/SEEDA-CECNSM49515.2020.9221795.

[70] W. Gul, M. Shams, and D. Al-Khalili, "FinFET 6T-SRAM all-digital Compute-in-Memory for artificial intelligence applications: An overview and analysis," *Micromachines*, vol. 14, no. 8, p. 1535, Jul. 2023.

[71] M. Harada, M. Takahashi, S. Sakai, and T. Morie, "A time-domain analog weighted-sum calculation circuit using ferroelectric-gate field-effect transistors for artificial intelligence processors," *Japanese J. Appl. Phys.*, vol. 59, no. 4, Apr. 2020, Art. no. 040604.

[72] A. Kiyonaga, "We need a taxonomy of working memory," *J. Cognition*, vol. 2, no. 1, Aug. 2019.

[73] A.-E.-H. Munir and W. M. Qazi, "Artificial subjectivity: Personal semantic memory model for cognitive agents," *Appl. Sci.*, vol. 12, no. 4, p. 1903, Feb. 2022.

[74] M. R. Schreiner, A. Bröder, and T. Meiser, "Agency effects on the binding of event elements in episodic memory," *Quart. J. Experim. Psychol.*, vol. 77, no. 6, pp. 1201–1220, Jun. 2024.

[75] M. Li and M. Moens, "Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering," in *Proc. Aaai Conf. Artif. Intell.*,] Jan. 2022.

[76] M. Badrouni, C. Katar, and W. Inoubli, "Large-scale knowledge graph representation learning," *Knowl. Inf. Syst.*, vol. 66, no. 9, pp. 5479–5499, Sep. 2024, doi: 10.1007/s10115-024-02131-5.

[77] L. Wang, W. Huang, Y. Li, J. Evans, and S. He, "Multi-AI competing and winning against humans in iterated rock-paper-scissors game," *Sci. Rep.*, vol. 10, no. 1, Aug. 2020.

[78] J. Tabaszewska, "Affective future and non-existent history: The issue of future past in memory research," *Memory Stud.*, vol. 16, no. 4, pp. 928–941, Aug. 2023.

[79] J. Chen, J. Xu, M. Bo, and H. Tang, "Augmenting embedding projection with entity descriptions for knowledge graph completion," *IEEE Access*, vol. 9, pp. 159955–159964, 2021.

[80] "Modeling of the technological process for metal production using a probability finite automatic machine," *Int. J. Innov. Technol. Exploring Eng.*, 2019.

[81] X. Ji, S. Hao, K. G. Lim, S. Zhong, and R. Zhao, "Artificial working memory constructed by planar 2D channel memristors enabling brain-inspired hierarchical memory systems," *Adv. Intell. Syst.*, vol. 4, no. 3, Mar. 2022.

[82] S. Singh and S. Seshadri, "Episodic question answering for cognitive agents," Tech. Rep., 2024, doi: 10.21203/rs.3.rs-4351479/v1.

[83] H. T. Maddali, E. Dixon, A. Pradhan, and A. Lazar, "Investigating the potential of artificial intelligence powered interfaces to support different types of memory for people with dementia," Tech. Rep., 2022.

[84] M. Wang, L. Lv, X. Xu, Y. Wang, Q. Ye, and J. Ni, "Navigable proximity graph-driven native hybrid queries with structured and unstructured constraints," Tech. Rep., 2022.

[85] T. Teofili and J. Lin, "Lucene for approximate nearest-neighbors search on arbitrary dense vectors," Tech. Rep., 2019.

[86] K. Echihabi, K. Zoumpatianos, T. Palpanas, and H. Benbrahim, "Return of the lernaean hydra: Experimental evaluation of data series approximate similarity search," *Proc. VLDB Endowment*, vol. 13, no. 3, pp. 403–420, Nov. 2019, doi: 10.14778/3368289.3368303.

[87] S. Kale, "From text to recommendations: How vector databases are revolutionizing personalized content delivery," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 4, pp. 3376–3387, Apr. 2024.

[88] W. Wang, Y. Wang, S. Joty, and S. C. H. Hoi, "RAP-gen: Retrieval-augmented patch generation with CodeT5 for automatic program repair," in *Proc. 31st ACM Joint Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, Nov. 2023, pp. 146–158, doi: 10.1145/3611643.3616256.

[89] H. Yang, "PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter," Tech. Rep., 2023.

[90] R. Zhao, H. Chen, W. Wang, F. Jiao, D. Long, C. Qin, B. Ding, X. Guo, M. Li, X. Li, and S. Joty, "Retrieving multimodal information for augmented generation: A survey," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, doi: 10.18653/v1/2023.findings-emnlp.314.

[91] M. Ramezani, M.-R. Feizi-Derakhshi, and M.-A. Balafar, "Knowledge graph-enabled text-based automatic personality prediction," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–18, Jun. 2022.

[92] C. Xie, L. Zhang, and Z. Zhong, "A novel method for constructing spatiotemporal knowledge graph for maritime ship activities," *Electronics*, vol. 12, no. 15, p. 3205, Jul. 2023.

[93] H. O. Ilhan, G. Serbes, and N. Aydin, "Decision and feature level fusion of deep features extracted from public COVID-19 data-sets," *Int. J. Speech Technol.*, vol. 52, no. 8, pp. 8551–8571, Jun. 2022.

[94] D. Zacharopoulou, A. Skopeliti, and B. Nakos, "Assessment and visualization of OSM consistency for European cities," *ISPRS Int. J. Geo-Information*, vol. 10, no. 6, p. 361, May 2021.

[95] J. Camenisch, M. Dubovitskaya, and A. Rial, "Concise UC zero-knowledge proofs for oblivious updatable databases," in *Proc. IEEE 34th Comput. Secur. Found. Symp. (CSF)*, Jun. 2021, pp. 1–16, doi: 10.1109/CSF51468.2021.00008.

[96] B. Shbita, C. A. Knoblock, W. Duan, Y.-Y. Chiang, J. H. Uhl, and S. Leyk, "Building spatio-temporal knowledge graphs from vectorized topographic historical maps," *Semantic Web*, vol. 14, no. 3, pp. 527–549, Apr. 2023.

[97] A. A. Wank, J. R. Andrews-Hanna, and M. D. Grilli, "Searching for the past: Exploring the dynamics of direct and generative autobiographical memory reconstruction among young and cognitively normal older adults," *Memory Cognition*, vol. 49, no. 3, pp. 422–437, Apr. 2021.

[98] V. Wardell, C. R. Madan, T. J. Jameson, C. M. Cocquyt, K. Checknita, H. Liu, and D. J. Palombo, "How emotion influences the details recalled in autobiographical memory," *Appl. Cognit. Psychol.*, vol. 35, no. 6, pp. 1454–1465, Nov. 2021, doi: 10.1002/acp.3877.

[99] B. Paneru, B. Thapa, and B. Paneru, "Leveraging AI in ayurvedic agriculture: A RAG chatbot for comprehensive medicinal plant insights using hybrid deep learning approaches," *Telematics Informat. Rep.*, vol. 16, Dec. 2024, Art. no. 100181, doi: 10.1016/j.teler.2024.100181.

[100] J. Pei et al., "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, Aug. 2019.

[101] O. I. Sheremet, O. V. Sadovoi, K. S. Sheremet, and Y. V. Sokhina, "Effective documentation practices for enhancing user interaction through GPT-powered conversational interfaces," *Appl. Aspects Inf. Technol.*, vol. 7, no. 2, pp. 135–150, May 2024.

[102] M. Kubica and D. Kania, "Technology mapping of FSM oriented to LUT-based FPGA," *Appl. Sci.*, vol. 10, no. 11, p. 3926, Jun. 2020.

[103] T. J. Gaffney, "Memory depth of finite state machine strategies for the iterated Prisoner's dilemma," 2019, *arXiv:1912.04493*.

[104] Y. Ahn, S.-G. Lee, J. Shim, and J. Park, "Retrieval-augmented response generation for knowledge-grounded conversation in the wild," *IEEE Access*, vol. 10, pp. 131374–131385, 2022.

[105] S. Ge, C. Xiong, C. Rosset, A. Overwijk, J. Han, and P. Bennett, "Augmenting zero-shot dense retrievers with plug-in Mixture-of-Memories," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 1796–1812, doi: 10.18653/v1/2023.emnlp-main.111.

[106] Z. J. Wang, "MeMemo: On-device retrieval augmentation for private and personalized text generation," 2024, *arXiv:2407.01972*.

[107] Z. Shao, "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy," 2023, *arXiv:2305.15294*.

[108] P.-H. Li, "Augmenting large language models with reverse proxy style retrieval augmented generation for higher factual accuracy," Tech. Rep., 2024.

[109] S. Lu, N. Duan, H. Han, D. Guo, S.-W. Hwang, and A. Svyatkovskiy, "ReACC: A retrieval-augmented code completion framework," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, doi: 10.18653/v1/2022.acl-long.431.

[110] E. M. Bir, Z. Kahn, J. Kalman, O. Rusz, M. Pski, I. Tam, G. Kelemen, R. Dus, G. Drtos, and C. Hamvai, "Cognitive functioning and psychological well-being in breast cancer patients on endocrine therapy," *Vivo*, vol. 33, no. 4, pp. 1381–1392, 2019.

[111] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proc. Aaai Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 17754–17762.

[112] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agricult.*, vol. 147, pp. 70–90, Apr. 2018.

[113] A. G. Ororbia and M. A. Kelly, "Cogngen: Constructing the kernel of a hyperdimensional predictive processing cognitive architecture," Tech. Rep., 2022, doi: 10.31234/osf.io/g6hf4.

NUR ARIFIN AKBAR (Member, IEEE) was born in Indonesia. He is currently pursuing the Ph.D. degree in mathematics and computer science with the University of Palermo, Italy, as part of the Marie Skłodowska-Curie Actions (MSCA) Doctoral Network.

He is a Doctoral Researcher at the University of Palermo, leading research in agentic large language model-based data sharing systems for agricultural environments.

Mr. Akbar is a member of several professional organizations and has been recognized internationally, including receiving the Silver Medal at the World Innovation Day (UN) Blockchain for Good Category and being named in the Top Ten Shell Livewire Energy Solutions. He has served as a reviewer for multiple Q1 and Q2 journals in artificial intelligence and computer science, including publications by Elsevier and Springer Nature.

RAHOOL DEMBANI is currently a Data Scientist and an Entrepreneur in Pakistan, with over eight years of experience in software development. He co-founded AgriDunya Technologies and is a Doctoral Researcher at the Marie Skłodowska-Curie Doctoral Network (Entrust) hosted by SingularLogic. His research focuses on addressing weak data verification and provenance issues in agriculture by developing privacy-preserving technologies. Leveraging MLOps practices, he ensures the reliable and efficient deployment of machine learning models to enhance data integrity and trust in the agri-data ecosystem. He has contributed to projects aimed at fostering sustainable and innovative agricultural practices through advanced computational solutions. His current research interests include federated learning, secure data sharing, and the application of AI-driven technologies to improve agricultural data management and decision-making processes.

BIAGIO LENZITTI received the M.Sc. degree in mathematics and applications from the Università degli Studi di Palermo, Palermo, Italy.

He is currently a Researcher with the Department of Mathematics and Informatics, Università degli Studi di Palermo. His academic work spans various fields, including health informatics and machine learning, with a focus on developing methodologies for improving access to medical information and enhancing patient empowerment. He has contributed to multiple international research projects and has authored numerous publications on topics such as autonomous navigation systems, e-learning object extraction, and health information systems. His current research interests include cyber risk assessment, distributed data sharing architectures, and tailored health information systems. His work integrates advanced computational techniques to address challenges in data security, artificial intelligence, and medical informatics.

DOMENICO TEGOLO was born in Italy. He received the M.Sc. degree in mathematics and applications from the Università degli Studi di Palermo, Palermo, Italy, in 1985.

He is currently an Associate Professor of computational vision and programming with the Department of Mathematics and Informatics, University of Palermo, where he previously was the Co-Director. He has dedicated his academic career to computational vision and advanced mathematical applications. He has held multiple prestigious research positions, including an Associate Researcher at the Institute of Biophysics, National Research Council (CNR-IBF), and a Key Member with the Human Brain Project Team. His professional experience also includes research roles at the National Institute for Astrophysics, Italian Institute for Nuclear Physics in Palermo, and ITALTEL Telecommunication SpA. He is an Associate Member at the Innovation Value Institute, Maynooth University, and has authored numerous research articles in computer vision and related fields. He is an Accomplished Researcher with extensive expertise in mathematical modeling of complex systems. His current research interests include biomedical image analysis, three-dimensional feature detection, and innovative parallel and serial architectures for image analysis. He has made significant contributions to fields including artificial intelligence, neural networks, and computational imaging techniques.