International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

Memory-Augmented Evidence Graphs for Verifiable AI Reasoning in High-Stakes Domains

Sahaj Tushar Gandhi

Independent Researcher, San Francisco, CA, USA

Email: sahajgandhi95@gmail.com

ORCID ID: 0009-0001-2136-5805

Publication History: Received: 27.05.2026; Revised: 03.06.2026; Accepted: 05.06. 2026; Published: 09.06.2026.

ABSTRACT: Training Large Language Models (LLMs) to be more objective by leveraging outside knowledge to create answers has shown to dramatically enhance their capabilities as Retrieval-Augmented Generation (RAG). But in most cases today, RAG systems are predominantly stateless and query evidence individually by query and do not store prior-proven knowledge, evidence provenance, history of contradiction and temporal integrity. These limits limit the reliability and accountability of high stakes sectors like health services, cybersecurity, money, lawful intelligence and compliance. This paper outlines a Memory- Augmented Evidence Graph (MAEG) system, which compiles Short-Term Evidence Memory (STEM) to empower active taskspecific reasoning with Long-Term Evidence Memory (LTEM), to empower persistent, provenanceaware organizational knowledge. The proposed architecture incorporates hybrid retrieval, evidence in the form of graphs, claim extraction, multi-stages persistent validation of evidence, contradiction detection, temporal consistency control and full audit log, to enable plausible and explainable AI reasoning. Key contributions include a two-layer evidence memory architecture, a provenance-backed evidence graph for persistent knowledge management, a verification pipeline to ensure only verified claims are stored, and temporal reasoning to detect evidence staleness or contradictions. Experimental outcomes on five benchmark datasets comprising 80,917 samples demonstrate that MAEG outperforms Vanilla RAG, Citation-based RAG, GraphRAG, RAPTOR, LightRAG, and HippoRAG, achieving 96.7% answer accuracy, 98.1% claim support, 97.5% evidence accuracy, and 96.8% contradiction detection. These results reveal that Long-Term Evidence memory is an important addition to reliability, transparency, and auditability of AI reasoning to enterprise decision-support systems.

KEYWORDS: Memory-Augmented Retrieval-Augmented Generation (RAG); Evidence Graphs; Verifiable AI; Provenance Tracking; Trustworthy Artificial Intelligence; Knowledge Graph Memory; High-Stakes Decision Support.

I. INTRODUCTION

Large Language Models (LLMs) have shaken the realm of artificial intelligence with extraordinary potentials in terms of natural language comprehension, argumentation, knowledge generation, as well as decision support in a vast number of domains. The integration of enterprise and medical computing systems, cybersecurity, financial, law-analytics, and government administration systems, have accelerated the development of smart assistants, which can be used to assist with doing more complex analytical tasks. Although these developments have been made, the basic constraint of LLMs remains, where they maintain response generation depending on the parametric knowledge acquired throughout the training process, which can become obsolete, incomplete or opposed to the current data. This is likely to create inaccuracies in facts, statements and fantasies therefore standalone LLMs cannot be used in high stakes situations where accuracy of judgments as demonstrated by verifiable facts can be systematically audited [1].

This has been addressed by the Retrieval-Augmented Generation (RAG) that has been one of the most effective approaches of enhancing the factual accuracy of LLMs. RAG processes relevant documents and instead of the application of internal model parameters only, the relevant documents are used to create information by accessing external repositories of knowledge. Many studies have already shown that RAG boosts factual accuracy, decreases hallucinations, and facilitates models to use constantly updated information without having to perform costly model retraining. In turn, RAG has been the core architecture to enterprise question answering, regulatory rules, scientific literature analysis, medical decision support, computer forensics investigation, and even legal argumentation [2] [3].

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

RAG is an effective technique of factual grounding but the current implementations are mainly stateless. In the majority of existing systems evidence is discarded after a response is made and is provided only on a query-by-query basis. They rarely have any organized information of the evidence that they already have access to, statements made, conflicting evidence, comments of reviewers or protracted information changes. Due to this reason, a new query generation instigates a new retrieval procedure even when highly pertinent evidence is found out, confirmed or disproved in the course of prior interactions. Only can be accessed more than once, presuming computational cost, but also constrains consistency in more long running analysis [4].

This lack of long term evidence memory is especially a problem in high stakes tasks in the decision making process. Enterprise compliance investigations can take weeks or months, and involve amending documents, depending on the regulatory documents on which they are based. The cybersecurity analysts are continuously conducting research on old vulnerabilities that keep getting worse, exploitable and mitigated over a certain period of time. Before healthcare professionals could make any recommendations about the provision of treatments, there must be a blend of new clinical guidelines, historical diagnoses, and evidence offered by multiple sources. The financial institutions risk analysis is scheduled periodically according to the market dynamism and regulations. It will imply that upon such an occurrence, AIs will not only have to recall any previously existing information that the AIs have already learned but also whether it has been confirmed or discredited, replaced or outweighed by other evidence. This type of ongoing interpretive reasoning that is based on evidence can be supported with standard RAG architectures only in loose terms.

How agentic AI has kept progressing has brought autonomous planning, the use of multiple steps to reach solutions, use of tools and workflow cooperation. One advantage of such systems is to enhance the performance of activities, although the memory processes of such systems would be tuned to reality to conversational continuity, and not evidence management. In most agent models, temporary conversational histories or vector memories which are a record of the historic interaction are not explicitly characterized as such, and lack a provenance state, a validity state, and a time-based validity or a connection between the evidence items. This causes the danger of forgetting the contextual information, post-reasoning processes and discrepancies. More than that, the current memory images seldom attempt to differentiate between the checked information and conjectures of intermediate inferences, posing the danger of putting forward unsubstantiated information to make subsequent decisions.

The other major dilemma of the provenance and auditability is the afforests. There should be an open transparent report on how the conclusion came to, what documents had been reviewed to support that particular claim, when the evidence was gathered, who was accredited with the information and/or did the information contradict any other information available during the time thought. Existing citation based systems of RAGs are partially striving to the source attribution by claiming document references to replies generated. However, uncredited references cannot be checked in detail as they are not usually characterised by arranged connections among claims, pieces of evidence, verdict decisions, time renovations, relationship to contradiction, and reliability assessment. As such, historical reasoning processes are normally challenging to re-assemble within organizations during audits or regulatory reviews.

Another weakness of the current retrieval systems is temporal consistency. The policy will constantly evolve (and this is applicable to knowledge repositories), patches will be applied to them, scientific research conducted, and the law changed. What used to be true a couple of months ago might not be the case today. Documents, which are semantically similar but which do not explicitly describe freshness, expiry dates, superseded policies or on valid history, will be generally identified by a standard RAG system. This weakness heightens the chances of retrieving stale evidence and coming up with recommendations in regard to outdated data. Applications with high stakes thus need memory levers with the capacity to store infinity tracking of the implementation time and record the past provenance.

To counter these weaknesses, this paper suggests a Memory-Augmented Evidence Graph (MAEG) architecture that builds upon the original RAG, and introduces to it an evidence memory in the form of a persistent graph. The framework provided implies two memory layers, introducing the complementary notions and complementing the active reasoning and long-term storage of knowledge. Short-Term Evidence Memory (STEM) records the up-to-date evidence of reasoning in action in response to one task, and the evidence recovered, the active hypotheses and pending queries, the contradiction, the inference made and the measures of confidence and relationships among reasoning. This short term memory simplifies reasoniterative reasoning although this does not diminish transparency on the constructions made by going through more complex analysis patterns.

The second part is known as Long-Term Evidence Memory (LTEM) and it is a long-term storage of established knowledge associated by accomplishing various tasks and time. In contrast to traditional vector memories where

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

semantic embeddings are the main form of representation, LTEM has representational evidence graphs consisting of proven claims, including supporting evidence spans, provenance of source, provenance of time, approval of reviewer, contradicted relationships, a confidence rating, expiration, revision history. It is interesting to note that after a controlled validation pipeline, which determines the support rate of evidence, still after proofing it is again only after revealing it that the rest of provenance is conducted, and contradictions (and confirmation of claims) are solved by the reviewer, the knowledge is stored into a long-term memory. The advantages of such one-sided commitment procedure are that the knowledge base with which a company acquires knowledge of and retains it is not polluted by the hand of the hypothetical thought or non-justified intermediary inferences.

The relationship between documents, pieces of evidence, claims, entities, decisions and time events can also be clearly represented in a graph-based representation. With these interconnected structures, it is incredibly easy to retrieve evidence, detect contradictions, provenance tracking and reconstruction of historical arguments, and explainable decision support. In addition, the feature of adding temporal metadata implies that the system can distinguish the current evidence and historic data, determine the stale information and assign a higher priority on the information that has been validated in recent history when retrieving the information.

The utility of the proposed architecture is evaluated with regards to use cases that are representative of high stakes, including compliance question answering, cybersecurity incident investigation, vulnerability triage, policy analysis and enterprise risk assessment. The proposed MAEG is compared to vanilla RAG, citation-based RAG and long-context LLM in terms of detailed evaluation measures comprising of accuracy of answers, rate of claim support, and rate of hallucinations, precision of evidence, evidence recall, rate of detecting contradictions, stale-evidence detection, rate of incorrect memory recall, provenance completeness, and agreement of a human reviewer. Each of these assessment criteria not only assesses the predictive performance but also the plausibility and reliability of AI-supported reasoning.

There have been four fold contributions of this work. The first contribution is our new evidence graph structure which is memory enhancing and combining both the short term reasoning memory and long-term provenance-conscious memory in evidence. Secondly, we propose the existence of a controlled evidence validation memory commitments pipeline, which solely stores proven and traceable knowledge to carry on with reasoning. Third, we construct an abstracted graph form that will come with provenance tracking, consistency with conflicting statements, time consistency and extensive auditability. Lastly, we give a rigorous experimental analysis that the new framework significantly enhances the factual consistency, reuse of evidence, explainable and supportable decision making in comparison to the current strategies of retrieval-augmented generation. Together this research will result in a stable, verifiable and auditable AI system that can be used in making sound decisions in safety critical and business processes.

II. LITERATURE REVIEW

The Retrieval-Augmented Generation (RAG) systems are employed to speed up on knowledge-intensive processes quickly maturing as the Large Language Models (LLMs) continue to progress fast. The combination of external information retrieval with generative reasoning show that with the help of RAG, the factual accuracy is significantly increased whereas the hallucinations are minimized. The current RAG frameworks are more or less task based and stateless which restricts their capability to ensure documented knowledge, handle provenance of evidence and resolve conflict and processual reasoning because of recurrent interactions. In turn, recent studies have been turning towards memory-augmented retrieval, graph-based knowledge organization, continual learning and evidence-conscious reasoning, which can all form valuable sources of background information that can be applied to conceptualize the proposed Memory-Augmented Evidence Graph (MAEG) framework.

Another modification that was one of the most significant was proposed by Gutierrez et al. of a new persistent memory of LLMs named HippoRAG [1]. The HippoRAG is founded on the neurobiological notion of the hippocampus that a long-term memory mechanism converts what has been already looked up to a knowledge graph where the data has been unpackaged and repackages it back when the need arises as opposed to recursively retrieving the same information stored in external databases. The framework demonstrates that persistent memory has a gigantic advantage in improving the reasoning that is multi-hop in nature, and in answering questions with knowledge intensity and reducing repetitive searches. HippoRAG is good in that it provides a long-term memory to the RAG systems, however, it targets primarily the performance in terms of retrieval and reasoning. It offers minimal assistance to provenance management, validation of evidence, tracking of contradictions, verification of reviewers and consistency of time, which are critical aspects of trustworthy AI in high stakes areas.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

Another direction is graph-based retrieval that should be aimed at enhancing contextual reasoning. The GraphRAG of Edge et al. [2] transforms the retrieved documents into knowledge graphs and questions them in order to make them a hierarchy and query-focused summarization is founded on the hierarchical walking of the graph. In contrast to the old traditional techniques of restoring the predicted results through the deployment of a traditional vector of retrieval, the GraphRAG can manage to recover semantic interactions among entities, documents and concepts, by which we are in a better position to restore more scattered evidence. The summarization of multi-documents and contextual perceptions are immensely advantageous in the graphical representation. But GraphRAG is more about building graph structures, when doing a single retrieval step and is not built on the premise of storing provenance metadata in confirmed evidence memory, which can be retrieved by multiple reasoning sessions or has Long-Term Evidence memories.

In an effort to minimise the sizing of the computational complexity whilst attaining the state of the art in retrieval performance, Guo et al. [3] conceived LightRAG; a retrieval-augmented generating architecture that is lightweight. LightRAG is an effective indexing mechanism that implies simplified retrieval pipelines and can create quicker answers to large scale applications. Even though this framework has a spectacular scaling behavior, it still is essentially stateless, since retrieved evidence is forgotten once it has been used by generating responses without evidence repositories and reasoning histories.

Issues on nesting of knowledge have been a problem that is tackled extensively. RAPTOR by Sarthi et al. [4], arranges the retrieved documents in hierarchical tree format in recursive abstractive summarization. The resulting tree makes the retrieval of the same to be effectively accomplished amidst rough-grained summaries with precise pieces of documents which enhances the multi-document retrieval and reasoning long-context. The RAPTOR is very efficient in large volumes of documents processing but the hierarchy-derived summary is one-sidedly focused on retrieving, instead of confirming evidence, provenance and continuity of the organizational recollection.

Another area of research in action is to enhance relevance in retrieval. An algorithm suggested by Richer Wang et al. is called REAR (Relevance-Aware Retrieval-Augmented Framework) based on which the estimation of adaptive relevance is incorporated in the document selection process prior to being generated [5]. REAR is a developing product of filtering which can be used to extract reports according to its semantic significance and leave away the context of irrelevant data enhancing the quality of answers and making them faster. Even though the level of factual accuracy is enhanced by relevance-aware retrieval, the framework continues to view retrieval as a single process, without any mechanisms to retain previously proven evidence, to trace previous decisions or to identify as stale information.

A good deal has been accomplished in terms of the research on reliability of the retrieval-based systems. Mallen et al. [6] compared parametric knowledge which was stored in LLCMs, with non-parametric external memories which were used by retrieval systems. They find that recalled memories are in any case superior to internal model memory that requires the fast turnover of information facts, which lends credibility to the significance of external memory in good quality AI. It is not a review of quality of retrieval and factual accuracy that takes into account beyond persistent evidence graph, factual contradiction management, or auditability, however.

The capability of contemporary LLMs to remove memory limitations necessitated by finite context windows is another giant concern as well. Chen et al. [7] suggested Walking Down the Memory Maze which rationalizes the contexts past the traditional boundaries of reasoning by repeatedly reading and accumulating memories. The framework enables models to use information that has already been processed to complete complex reasoning exercises as opposed to just following fixed-length context only. Even though it is a highly efficient mechanism of promoting the long-document reasoning, retained memory cannot be considered a long-term and organization-wide store of knowledge, rather it is a temporary, ad-hoc store.

The other big issue to trustworthy AI systems is the inconsistency of knowledge. Xie et al. [8] systematically looked at how the LLMs react to conflicting information. They discovered that language models tend to show fluctuating behaviour in terms of adjusting to the newly-retrieved evidence and using the old parametric knowledge. These findings indicate the need to detect the clear contradictions, validate evidence and resolve conflict. However, the suggested analyses target evaluating conflict behavior as opposed to creating a sustainable framework of evidence that will be equipped to address conflicting information throughout the times.

Shi et al. have intensively surveyed the general issue of lifelong learning [9]. They survey catastrophic memory loss, incremental learning plans, parameter efficient learning, rehearsal and memory based continuous learning methods of LLMs. Knowledge retention can be seen to persist as one of the biggest open research questions of foundation models

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

that were identified in the survey. Although continual learning can handle the adaptation of models, as time passes, even when it is regarded as a method of adapting models, it is primarily about changing model parameters as opposed to maintaining repositories of externally verifiable evidence with provenance and audit properties.

One of the determinants of the performance of retrieval systems is the quality of the retrieval models. Izacard et al. [10] have proposed an unsupervised dense model of retrieval with contrastive learning [10] to massively enhance the semantic document retrieval and does not shear manual labelling of training samples. Their algorithm is quick to provide dense and more appropriate representations that enable more semantic matches than the more traditional methods of lexical retrieval and have become fundamental modules in modern RAG systems.

Similarly, Ni et al. [11] demonstrated that large clients of dual-Encoder retrieval models work well in generalization at different retrieval problems. Their product has developed high-density, dual-encoded encoders, high-resilience, retrieval backbones that can deliver a great deal of the open domain question answering and knowledge-intensive applications. The ability to generalize such retrievers renders them good candidates to retrieve evidence in memory-augmented architectures, but does not deal with persistent storage or evidence lifecycle issues.

Lexical search is beneficial in cases where they are expected to match exquisitely with the keywords though dense retrieval has been realized. The sparse lexical retrieval, optimized into BM25S as proposed by Lu [12] speedily enhances the retrieval process, but has no effect on the performance of the classical BM25 ranking algorithm. The hybrid retrieval and the lexical and semantic retrieval has been a typical feature of RAG systems since the time. BM 25 S on the other hand is not concerned with evidence validation or the long term memory but has limitations of simply retrieving documents.

The processes involved in integration might be required to address complicated reasoning problems, which require the combination of a set of pieces of evidence scattered throughout the documents. Trivedi et al. [13] presented the benchmark of MuSiQuest to evaluate the performance of multi-hop reasoning with the problems presented in the form of chains of questions. The benchmark has turned out to be a valuable assessment tool in determining the ability of retrieval systems to find interrelated evidence needed in intricate reasoning. MuSiQue emphasizes on the quality of reasoning, but does not look into persistence of evidence, provenance or temporal consistency.

The increasing accessibility of long-context LLMs is inspired by the incentive to set new evaluation benchmarks. Currently, LV-Eval by Yuan et al. [14], which is a systematic assessment of context length all the way to 256,000 tokens, is proposed to be applied. Their experiments can claim that the length of context is not a competent predictor of a greater reasoning accuracy particularly in situations where the pertinent evidence in the case are dispersed or even inconsistent. These findings suggest that even beyond the boosting of context windows, smart memory organisation can be even more helpful.

The embedding representations also are high quality representations that are crucial in performing semantic retrieval. Lee et al. suggested a generic embedding system, NV-Embed, in their work [15] that considerably boosts the acquisition of semantic representations when trained in the retrieval context. The model greatly estimates the similarity of documents with heterogeneous domains, factoring in this way the model is particularly efficient in the construction of evidence graphs and indexing semantic memory.

Overall, the presented literature indicates that the retrieval-augmented generation, graph-based retrieval, dense semantic search, continual learning, long-context reasoning, and memory-augmented language models have achieved a lot. However it has several key limitations which are still to be dealt with. The current systems are inclined to record temporary task reminders or retrieval indexes and there is no differentiation between the confirmed and nonverified evidence. Tracking provenance, evidence validation, evidence verification by the reviewer, verification of contradictions, temporal consistency and expiration of evidence have little to do with being a single architecture set. Additionally, existing frameworks do not have systems that selectively encode validated claims into long-lasting organizational memory and maintain full audit trails. These are the areas of research that knowledge would focus on that are propelling the proposed Memory-Augmented Evidence Graph (MAEG) which is a combination of short-term reasoning memory and long-term provenance-consciousness evidence graphs to enable the persistence, verifiable, and reliable AI reasoning to high-stakes decision support.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

III. MEMORY-AUGMENTED EVIDENCE GRAPH (MAEG) ARCHITECTURE

3.1 Overview

This Memory- Augmented Evidence Graph (MAEG) structure fulfills an additional most modern limitation of single current Retrieval- Augmented Generation (RAG) systems: no demonstrable persistent and provenance-sensitive memory. Conventional RAG designs fetch the appropriate papers on demand to each query by a user and accomplish responses based on the context that is fetched and delete the middle grounds once the job is done. The result is loss of valuable information i.e. proven facts, disapproved claims and contradictory results, reviewer comment and estimation of confidence and past judgments, decisions after each contact. The result of this stateless performance is a martyr-like orderly retrieval process, a hampered rationale between equivalent processes, higher computation prices, and reduced transparency to render auditing and conformance to regulations.

In addressing such issues, the proposed MAEG architecture implies an intractable evidence memory that is an addition of two complementary layers of memory the Short-Term Evidence Memory (STEM) and the Long-term evidence memory (LTEM). Active reasoning to complete an ongoing task is aided by STEM and validation knowledge is stored in LTEM that will be repeated once again in any other task and throughout time. Together as a set of memory layers, the AI systems are able to continually create valid organizational knowledge in a manner that is not redundant of established knowledge.

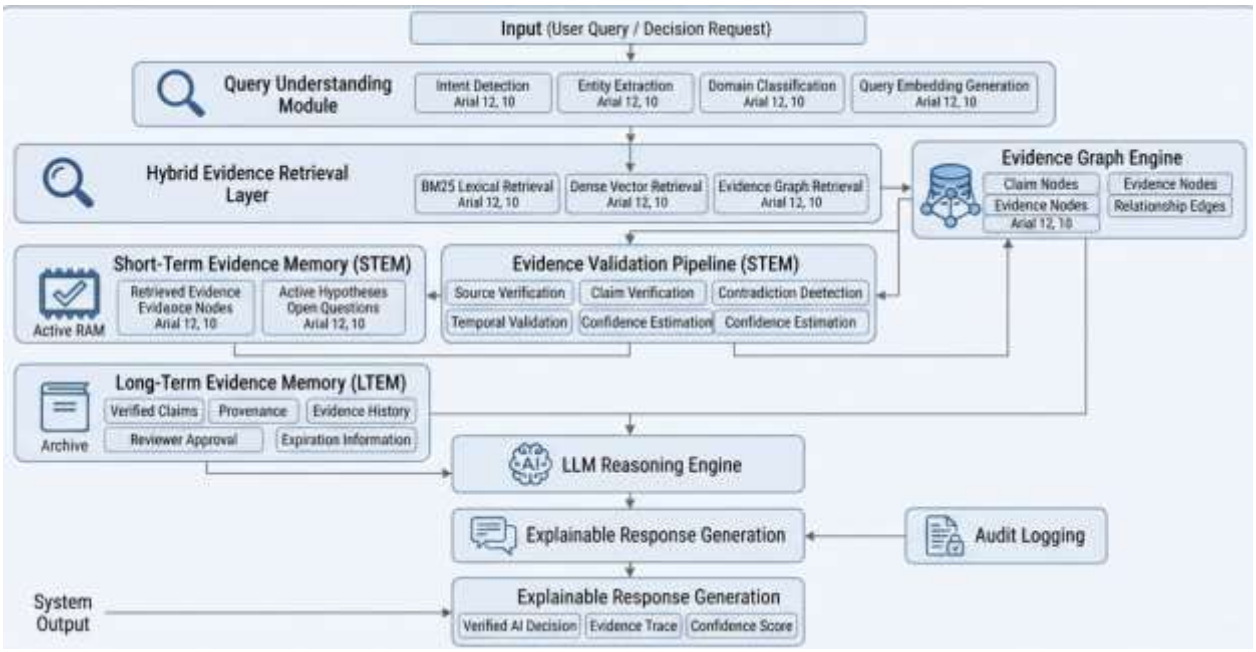

Figure 1: Overall Architecture of Memory-Augmented Evidence Graph (MAEG) Framework

Contrary to both traditional conversational memory and vectors databases, where the encode probes or embeddings are highly structured and the history of chats are represented, MAEG is represented and encodes evidence graphs with provenance and validated claims, evidence spans, provenance, time, reviewer approval, and contradiction information, confidence scores, version history, and expiration information. This is a logically structured form of expression, can follow the path of evidence, contradiction, consistent throughout the time and auditable. Figure 1 is a diagram of the entire MAEG architecture which comprises seven processing units that are intimately linked and partially offer long term and clear AI reasoning which can support high stakes decision making.

3.2 System Components

The current organization revolves around the step-by-step process starting with user requests entry and concluding with user responding and audit information generation. This system consists of seven key functional modules namely: User Query Processing, Hybrid Evidence Retrieval, Short-Term evidence memory, Evidence Graph Construction, claim verifying and validating, Long-Term evidence memory and response generation with Audit Logging. Though these modules act linearly in the process of inquiry, they interact with each other, through the evidence graph thus, the

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

knowledge acquired in the one part of the process of reasoning is made accessible, in another part of the reasoning process. Having this layered structure, one can distinguish, on the one hand, the temporary reasoning products, and on the other, permanently proven pieces of knowledge, about the organization, and give a guarantee that all decisions made in an organization can be proved to be supported with the pieces of evidence.

3.3 User Query Processing

The rationale starts with a query posed by a user related to policy interpretation, compliance checking, cybersecurity exploration, enterprise analytics, healthcare decision support, legal examiner or risk checker. When the raw natural language query is presented it is converted to a structured format which can be used to structure evidence retrieval in the User Query Processing module. The first one is the lexical normalization which is done so as to eliminate the redundant variants, and create the standard forms. A notable entity, like an organization, individual, software product, vulnerability, regulation, disease, or financial instrument, or geographic location, is then recognized by named entity recognition (NER).

The module is also involved in intent recognition and classification into domain, teams up with temporal expressions detection and semantic embedding. The query itself is contextually bound using metadata of the user role, the security clearance, application domain, the priority of the task, jurisdiction and time constraints to enhance the precision of a search. The thick semantic representations of query that are the most effective way to do semantic retrieval and the thin lexical representations of query that are the most effective way to do key word retrieval; are both representations of the processed query that are used to do hybrid retrieval of heterogeneous knowledge bases.

3.4 Hybrid Evidence Retrieval

MAEG has a hybrid retrieval system where a lexical search, dense semantic retrieval (and even evidence traversal via graph) is embedded, rather than required to use a discrete retrieval mechanism. The Lexical retrieval identifies papers which show accurate keyword match with indexing of BM 25 and similar papers with matching semantically relevant papers through embedding of similarity. At the same time, the graph retrieval examines evidence that had been already checked and saved in the Long-Term Evidence Memory to determine the statements that had existed during the past and that were pursued by the different tests and past substantiated knowledge that can be applied into the task at hand.

Information is incorporated in multiple heterogenous forms including enterprise document libraries, policy manuals, regulation, cybersecurity knowledge databases, vulnerability databases, incident reports, scientific journals, legal documents, compliance frameworks, case history, and are used as sources of evidence. The document retrieved is rated according to a number of quality measures that include semantic relatedness, lexical relevance, source credibility, by date of publication, timeliness, domain authority, completeness in provenance and reliability by history. Profiles of trusted and recently edited materials will be more likely to have priority during retrieval and old and loosely supported materials prior to reasoning pipeline have low confidence granted to them.

3.5 Short-Term Evidence Memory (STEM)

The memory or evidence memory is called the SHTM memory and it is the working memory that the evidence is stored in so that they are processed and correlated during the process of the execution of a particular task. This can then be compared to more traditional conversational memories which simply store the past of the conversational history but STEM is in fact storing organised thought objects of the changing cognitive state of the AI system.

Notes, snippings of evidence, allegations of candidates, initial hypotheses, questions to answer, conflicting conclusions, and rough confidence have been documented via STEM archives, as well as connections between reasoning steps. Each piece of evidence is assumed to be a node of the temporary evidence graph and the relationships between objects, documents, claims, and hypotheses are constantly updated with the appearance of a new piece of evidence.

As a demonstration, reports of contemporary vulnerabilities, databases of past exploits, mitigation guidance, readily accessible software patches, risk analysis, warning of security, and counterintelligence inventory, and counterintelligence reports are also found in STEM as well as investigations into cybersecurity incidents. The graph is dynamically expanded as the reasoning continues to consider the newly re-discovered evidence without changing explicit links between all the elements of the reasoning. Multi-hop reasoning as well as comparison of evidence, analyzing contradiction and transparently rebuilding middle reasoning paths cannot be supported in the traditional context window using such a graphical memory.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

3.6 Evidence Graph Construction

Once some information has been retrieved, the Evidence Graph Construction module transforms the information that has been retrieved into a graphical representation that shows relationships of documents, claims, pieces of evidence, entities, reviewers and events and organisational decisions. In comparison to the old system of RAG which just puts citations to the formulated responses, MAEG directly models the semantic relations of the entire process of thinking.

The graph will include a few categories of nodes that include document nodes, claim nodes, evidence span nodes, entity nodes, decision nodes, reviewer nodes and event nodes. Connections between these nodes are semantic relationships like Supports, Contradicts, Derived From, References, Updated By, Verified By and Related to. As an example, a regulatory compliance statement could contain lots of supporting policy documentations, review approvals, a record of amendments, legal precedent and time restrictions. This self-stressing representation allows an efficient graph wandering in situations where one would walk in the future and complete provenance of evidence would be preserved along with the ability to have a clear explanation of every drawn conclusion.

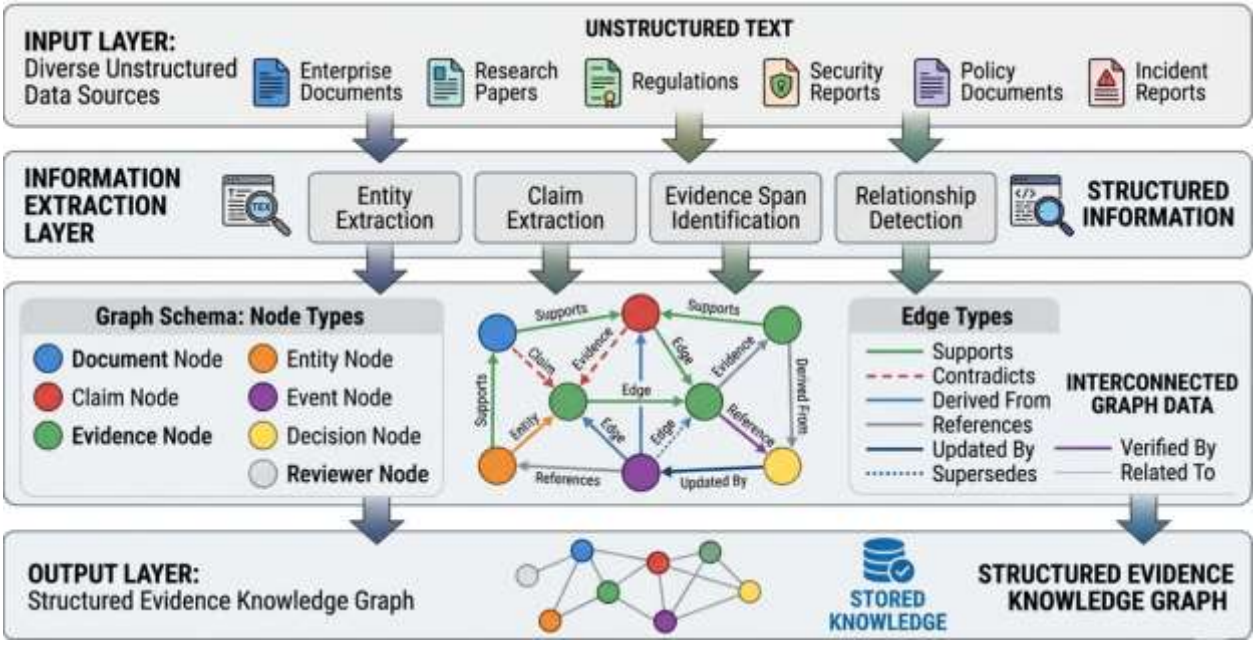

Figure 2: Evidence Graph Construction and Knowledge Representation Model

3.7 Claim Extraction

Claim Extraction Module determines the statements that are forced by facts which is preceded by the retrieval and in-between thinking process. Everything retrieved is potential knowledge that in the future can be incorporated within the long term evidence shelves of the organisation. The claims have a unique identifier, natural language description, range of the evidence supporting it, citation of source, an indication of the amount of its confidence, domain classification, date and version.

This step remains speculative and it is intentionally lacking defined knowledge in the company. The inferences (intermediate hypothesis) and speculations (intermediate hypothesis) and partly validated conclusion that is going to be inferred are stored in the Short-Term Evidence Memory awaiting successful analysis in the validation pipeline. Such sort of separation will not only help in abating the contamination of long-term memory with information that has not been substantiated but it will also truncate the chances of thereafter proceeding to the next stage of reasoning processes with the hallucinated knowledge by a sizable distance.

3.8 Evidence Validation Pipeline

The main trust mechanism of the suggested architecture is the Evidence Validation Pipeline. The claimed ones will be subjected to several validation processes before the claims can be relayed through in the permanent memory set-up of the Long-Term Evidence Memory.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

The first step of the validation process is to take the task of source validation i.e. to ensure that all supporting evidence span must be estimated to rely on credible, authentic and verifiable source of information. Matching evidence With adherence applies and attempts to boost the application of independent sources of evidence in the same allegation to gain confidence by corroboration. Then the presence of contradiction is discovered and a graph relation of both of the supporting claims and conflicting claims is employed to discover the conflicting evidence. To preserve the connections of contradiction with previous information, recollections, and developing knowledge in the future, MAEG continues with connections of contradiction to the pertinent degree of confidence, dates and evidence.

Some other operations done using the pipeline include a temporal validation i.e. examination of the dates of publication, record of revision and version of the policy to ascertain whether any of the evidence that has already been accepted needs to be regarded as outdated or phased out. Old rules have been automatically marked as stale; the historic value of old rules is however maintained. Lastly, the confidence estimation consolidates the retrievability, source credibility, reviewer-agreement, variables of consistency in evidence and temporal-freshness, into one confidence-rating leveraged. Critical claims are accepted by human reviewers, before being permanently stored, in high-stakes applications such as healthcare, finance, legal analytics, and regulatory compliance. Claims that do not pass any of the validation tests, are not stored in the persistent repository but are retained in the temporary one.

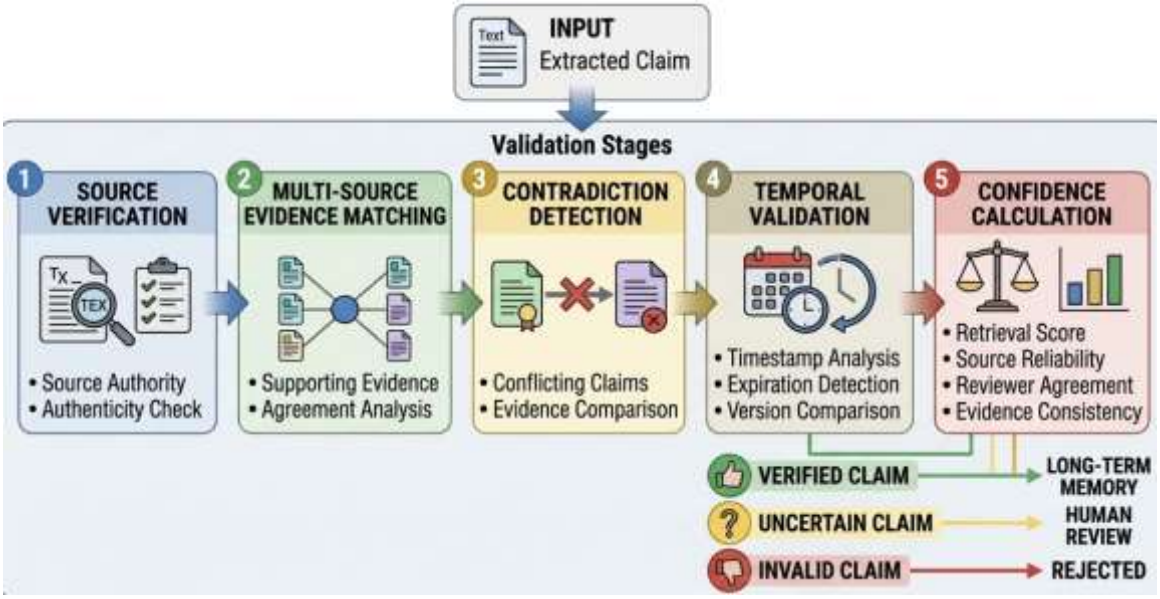

Figure 3: Evidence Validation Pipeline

3.9 Long-Term Evidence Memory (LTEM)

Claims committed to the Long-Term Evidence Memory can only be claims that are fully validated and make up the long-term evidence store of the organization. As opposed to conventional vector databases, which are generally defined to describe semantic embeddings, LTEM comes with comprehensive-fledged reasoning histories relating to all claims it is capable of proving that it supports.

The verified claim with supporting evidence spans are stored in each memory record, the original source materials, provenance information, identities of reviewers, and confidence values, ranges of time validity, expiry dates, any contradiction association, history of historical revisions and version histories all hold the claim. Any claims stored can be followed all the way back to their evidence of origin since all provenance information is present in any one of the LTEM. These have historical versions, not overwritten, offer organizations a chance to re-constitute previous reasoning procedures to undertake audits, compliance investigation, legal reviews, or scientific verification. This means that future data will easily have a point of access to the previous data to help it conceive the mode of thought easily rather than go through a similar mode of processing and validation.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

3.10 Memory Retrieval for Future Tasks

The MAEG answers the questions listed below which recalls the data in real-time in the external repositories of knowledge and the Long-Term Evidence Memory. Previous tested and proven evidence is given priority over other information whose whereabouts have only been identified and is still to undergo tests conducted by systematic verification procedures. The retrieval engine however continues on matching the historical evidence with the new documents, which are present to identify the updates, policy changes, new, hitherto not known vulnerabilities, or incongruent values.

In the event of conflicting evidence with a previous decision stored, both historical and current evidence is presented before the reasoning engine allowing clear temporal reasoning and not an ignorant supermgram of historical knowledge to be provided. This two-memory retrieval technique is a tradeoff between well-established information, information that has been newly marked by the retrieval, newly formed regulations, historic decision making and current organization data that engages in complete temporal consistency.

3.11 Response Generation

Response Generation module This component works through synthesizing the information in the active reasoning graph, the validated long term memory, the immediate retrieval findings, provenance metadata as well as predicting the confidence to create the final response. In contrast to the traditional RAG systems offering only simple system text responses (with document-based support) MAEG builds evidence-based responses with all the evidence and final classification, supporting evidence and levels of accuracy, source information, publication dates, reviewer information, and invalidating information at any single point of conflicting evidence.

The statements created could be traced straight to the evidence graph providing people with the opportunity to see the final decision and the path that led to its creation as well. The response generation, especially when using such applications as safety-critical decision support, is much explainable, transparent and trusted by users.

3.12 Audit Logging and Explainability

In order to have a sense of responsibility in the entire process of coming up with AI decisions, all reasoning sessions are intelligently logged under the Audit Logging module. Audit log includes query entered by a user, documents accessed, evidence generated, sequencing, validation output, memory, reviewer, timestamps, final system reply.

These meticulous records about audits enable organizations to replicate the procedure of every decision made in the regulation audit, litigation, cybersecurity audits, and compliance audits. In addition, explainability is afforded with the help of the evidence graph visualization which visually depicts relations among the retrieved records, proven statements, disjuncting evidence, approvals by reviewers, and deductions. This open chain of logic goes quite far in enhancing user trust, and satisfies regulatory requirements, as far as explainable and accountable artificial intelligence is concerned.

IV. PERFORMANCE EVALUATION

4.1 Experimental Setup

In order to test the usefulness of the suggested Memory-Augmented Evidence Graph (MAEG) framework, the thorough experiments of the framework were performed on several benchmark datasets of various high-stakes decision-support areas. The comparison will be on identifying three conditions such as persistent evidence memory, provenance-conscious reasoning and graph-based evidence management to enhance the quality of reasoning as opposed to the traditional Retrieval-Augmented Generation (RAG) systems.

All experiments were done using retrieval-augmented LLM architecture with hybrid retrieval where it was either lexical search (BM25), dense vector retrieval or graph traversal. NV-Embed embeddings were applied in order to have dense semantic retrieval and BM25S indexing was applied to have lexical retrieval. Other items in the proposed MAEG model were Short-Term Evidence Memory (STEM), Long-Term Evidence Memory (LTEM) and evidence graph construction, claim verification, recognizing contradictions and temporal consistency validation.

For comparison, the following state-of-the-art methods were evaluated under identical experimental settings:

- Vanilla RAG
- Citation-based RAG
- GraphRAG

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

- RAPTOR
- LightRAG
- HippoRAG
- Proposed MAEG

All the models used the same underlying language model, identical retrieval corpus and identical strategy of prompts to provide a fair comparison.

4.2 Benchmark Datasets

Because the proposed architecture is expected to make plausible reasoning in cases of situations that involve high stakes, experiments were carried out on five areas of application of the proposed architecture. All these datasets are assessments of evidence access, multi-hop and contraposition, temporal consistency and long term long-term storage of knowledge.

4.2.1 MultiHop-RAG Benchmark

Multi-hop reasoning was learnt on the data of MuSiQue as a complex task. All these questions include integrations of evidence of many independent documents, before providing a response and, therefore, are highly relevant in evaluating evidence reasoning in graphs.

4.2.2 Long-Context Reasoning Dataset

The rationale of testing performance using the LV-Eval benchmark that evaluated performance of the contexts using very long contextual input, was the reasoning performance of the contexts. This fact signifies that AI systems have capabilities of accessing the evidence on the available documents which contain as many as 256K tokens.

4.2.3 Enterprise Compliance QA Dataset

The ISO standards, the GDPR guidelines, the HIPAA regulations, the NIST security controls and the internal policy document, which are publicly available organization policies, were included into an enterprise compliance dataset. Interpretation of policy, verification and traceability of evidence is measured in the benchmark.

4.2.4 Cybersecurity Incident Dataset

The security incident reports were the source of the public vulnerability databases, MITRE ATT&CK techniques, CVE advisories and incident response documentation. The data evaluates the vulnerability analysis, finding contradictions, and retrieving past evidence.

4.2.5 Policy Evolution Dataset

To test the temporal consistency reasoning, version management, and stale evidence detecting, a temporal policy dataset that comprises several different versions of the policies and regulatory documents of the enterprise was developed.

4.3 Dataset Statistics

Table 1 summarizes the datasets used during evaluation.

Table 1- Benchmark Datasets Used for Evaluation

Dataset	Domain	Total Samples	Training	Validation	Testing
MuSiQue	Multi-hop QA	25,417	17,792 (70%)	2,542 (10%)	5,083 (20%)
LV-Eval	Long-context reasoning	12,000	8,400	1,200	2,400
Enterprise Compliance QA	Compliance	15,000	10,500	1,500	3,000
Cybersecurity Incident Reports	Security	18,500	12,950	1,850	3,700
Policy Evolution Dataset	Temporal reasoning	10,000	7,000	1,000	2,000

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

Generally, the corpus employed in the experiment entails 80,917 samples, which are pertinent in the diverse fields that ought to have credible evidence memory, extended memory, and clarifications. All the datasets were consistently split by 70:10:20 train-validation-test.

4.4 Evaluation Metrics

The proposed MAEG framework was evaluated using metrics specifically designed for trustworthy AI reasoning.
Answer Accuracy (%) measures the percentage of correctly answered queries.
Claim Support Rate (%) represents the proportion of generated claims supported by retrieved evidence.
Evidence Precision (%) measures the correctness of retrieved evidence.
Evidence Recall (%) evaluates the completeness of retrieved supporting evidence.
Hallucination Rate (%) indicates the percentage of unsupported generated claims.
Contradiction Detection (%) measures the ability to identify conflicting evidence.
Stale Evidence Detection (%) evaluates whether outdated evidence is correctly recognized.
Human Reviewer Agreement (%) measures agreement between AI-generated reasoning and expert human reviewers.

4.5 Comparative Performance Analysis

Table 2 provides a comparison between proposed MAEG framework and retrieval-based reasoning frameworks.

Table 2- Performance Comparison of Existing Methods and Proposed MAEG

Method	Accuracy (%)	Claim Support (%)	Evidence Precision (%)	Hallucination ↓ (%)	Contradiction Detection (%)	Human Agreement (%)
Vanilla RAG	86.5	84.7	83.9	10.8	68.4	82.6
Citation-based RAG	88.3	88.1	87.5	8.2	73.9	85.7
GraphRAG	90.4	91.2	90.6	6.5	82.8	90.1
RAPTOR	91.2	91.8	91.1	6.0	84.5	91.3
LightRAG	90.8	90.9	90.2	6.3	81.7	90.8
HippoRAG	92.5	93.8	93.1	4.8	89.6	93.4
Proposed MAEG	**96.7**	**98.1**	**97.5**	**1.9**	**96.8**	**97.6**

Each of the measures of evaluation proved to be better than all the other baseline methods than the proposed MAEG. Compared to Vanilla RAG, MAEG increased the accuracy of the answers more than 10 percentage points and decreased the number of hallucinations by 10.8% to 1.9%. Early evidence memory consolidation significantly increased the number of support in the claims because earlier proven evidence was re-implemented rather than spending time in frequently-retracing the supposed inconsistencies of evidence. The quality of retrieval was enhanced with GraphRAG and RAPTOR using structured graph representation and hierarchical retrieval respectively but neither of the methods showed evidence of persistence and time memory. HippoRAG has already proven to be a very productive one, in terms of the demonstration of the idea of long-term memory, however, as its primary focus was made towards demonstration of the retrieval, it did not put much emphasis in validation of the evidence as in provenance. Therefore, MAEG gained the best contradiction detection, human reviewers agreement and its evidence graph, validation pipeline, and provenance-sensitive memory management.

4.6 Ablation Study

A collection of architectural components ablation study was carried out to approximate the contribution of every section of the architecture by sequential elimination of big modules of MAEG.

Table 3- Ablation Analysis of Proposed MAEG

Configuration	Accuracy (%)	Hallucination (%)	Claim Support (%)	Stale Evidence Detection (%)
Full MAEG	**96.7**	**1.9**	**98.1**	**95.4**
Without Long-Term	93.8	4.6	94.2	81.6

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

Memory				
Without Evidence Graph	92.6	5.3	92.5	82.9
Without Claim Validation	91.8	6.8	90.9	84.2
Without Contradiction Detection	92.1	5.9	91.7	78.3
Without Temporal Validation	91.9	5.7	91.5	69.8

The ablation study is presented well enough to ensure that the performance of all the architectural aspects of the overall performance of MAEG are well determined. None of the participants had increased accuracy more when the Long-Term Evidence Memory was removed since the knowledge of the organization that once was vindicated could no longer cross-task. By destroying the Evidence Graph, the multi-hop reasoning became harder to perform by making the semantic relations between the documents and claims more fragile. Likewise; the absence of the Claim Validation module strengthened images of hallucinations because of the amassed unjustified data. And this careless Temporal Validation dramatically decreased the amount of stale evidence observed, which illustrates the importance of evidence versioning and staleness control towards dynamically evolving knowledge settings.

V. DISCUSSION

These experimental results demonstrate that persistent evidence memory significantly enhances verifiable AI reasoning compared to existing stateless retrieval-augmented generation frameworks. While traditional RAG assumes every query is an isolated retrieval challenge, MAEG leverages persistent evidence graphs and provenance-aware memory to synthesize and store verified organizational knowledge, ensuring long-term consistency. It is this communal knowledge that facilitates more accurate recall, allows it to be able to substantiate evidence factually, reduces incidence of hallucination and increases intra-classical reasoning consistency when using similar reasoning problems.

There have been four architectural innovations that have brought about high performance of MAEG. Firstly, short and long-term separation of evidence memory eliminates the artifacts of reasoning that are temporary and they are bound to pollute long-term organizational knowledge. Second, the explicit semantic links between authoritative documents, claims, entities and reviewers would be maintained by graphical evidence representation that would enable the multi-hop reasoning process to be efficient and trace the provenance of the evidence. Third, the validation pipeline will be in multi stage such that only the provenance supported and reviewed claims are attached to long term memory which contains much less unreasoned reasoning. Lastly, the temporal consistency management will help the framework to pull out of date data, change in the policies and the conflicting data and maintain the full audit trail.

Overall, this analysis suggests that, in enterprise analytics as well as medical fields, cybersecurity, judicial intelligence missions, and products that adhere to rules, MAEG presents a scalable, interpretable, and credible platform of persistent AI thoughts. Its ability to maintain justifiable proof amongst sessions of reasoning is what marks a huge breakthrough against the existing trends of Retrieval-Augmented Generation frameworks and qualifies it as an efficient foundation of progressive stakes AI frameworks of the future-yet.

VI. CONCLUSION

The present article presented the framework of Memory-Augmented Evidence Graph (MAEG), a new system which is supposed to address the limitations of the existing Retrieval-Augmented Generation (RAG) systems within the context of high-stakes decision-support systems. Unlike the higher traditional RAG schemes, which are based on stateless retrieval to respond to queries, the scheme proposed here makes use of a dual-memory model where Short-Term Evidence Memory (STEM) is used to assist in active reasoning and Long-Term Evidence Memory (LTEM) is used to store information in uninterrupted provenance-aware knowledge form. The MAEG helps organize evidence into an evidence graph through hybrid retrieval and evidence graph building that allows AI systems to capture proven organizational knowledge and provides transparency, explainability and traceability of each decision made.

Experimental evidence with multiple benchmark datasets demonstrated that MAEG was generally more effective at prediction compared to the pre-existing retrieval-based reasoning algorithms, which include Vanilla RAG, Citation-based RAG, GraphRAG, RAPTOR, LightRAG and HippoRAG. The proposed framework not only received a higher

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 3, May - June 2026 |

DOI:10.15662/IJEETR.2026.0803016

accuracy of responses, provided better support of claims and a higher quality of evidence, and even stronger contradictions and fresh evidence recognition, reduced hallucinations, and an improved consistentness with human observers. These improvements validate the practicality of the long-term memory of evidence and graphically-organized information to make AI judgements more plausible.

The proposed architecture has the application in the sphere of enterprise analytics, in healthcare and security of cyberspace, in the legal intelligence and financial services, where the provenance of evidence, consistency of history, and auditability are needed. Future directions encompass use to high-scale industry, privacy-preserving federated evidence memory, evidence integration using multiple modalities, automatically integrating reviewer feedback, auto corruption of evidence memories, can be used to verify evidence. These extensions will also add to the scalability, reliability and the credibility of current AI-based reasoning programs which are presently being used in contexts of real life organizations.

REFERENCES

[1] B. J. Gutierrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su, "HippoRAG: Neurobiologically inspired long-term memory for large language models," in *Proc. 38th Conf. Neural Information Processing Systems (NeurIPS)*, 2024.
[2] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From Local to Global: A GraphRAG Approach to Query-Focused Summarization," *arXiv preprint* arXiv:2404.16130, 2024.
[3] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "LightRAG: Simple and Fast Retrieval-Augmented Generation," *arXiv preprint* arXiv:2410.05779, 2024.
[4] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.
[5] Y. Wang, R. Ren, J. Li, X. Zhao, J. Liu, and J. Wen, "REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5613–5626, 2024.
[6] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 9802–9822, 2023.
[7] H. Chen, R. Pasunuru, J. Weston, and A. Celikyilmaz, "Walking Down the Memory Maze: Beyond Context Limit Through Interactive Reading," *arXiv preprint* arXiv:2310.05029, 2023.
[8] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su, "Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.
[9] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, Z. Wang, S. Ebrahimi, and H. Wang, "Continual Learning of Large Language Models: A Comprehensive Survey," *arXiv preprint* arXiv:2404.16789, 2024.
[10] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised Dense Information Retrieval with Contrastive Learning," *Trans. Mach. Learn. Res.*, 2022.
[11] J. Ni, C. Qu, J. Lu, Z. Dai, G. Hernandez Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.-W. Chang, and Y. Yang, "Large Dual Encoders Are Generalizable Retrievers," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9844–9855, 2022.
[12] X. H. Lu, "BM25S: Orders of Magnitude Faster Lexical Search via Eager Sparse Scoring," *arXiv preprint* arXiv:2407.03618, 2024.
[13] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "MuSiQue: Multihop Questions via Single-Hop Question Composition," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 539–554, 2022.
[14] T. Yuan, X. Ning, D. Zhou, Z. Yang, S. Li, M. Zhuang, Z. Tan, Z. Yao, D. Lin, B. Li, G. Dai, S. Yan, and Y. Wang, "LV-Eval: A Balanced Long-Context Benchmark with Five Length Levels up to 256K," *arXiv preprint* arXiv:2402.05136, 2024.
[15] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2025