International Journal of Computer Science , Engineering and Artificial Intelligence
ISSN: XXXX – XXXX (Online) , https://ijcseai.com/
Regular Issue, Vol. 2, Issue. 4 (October – December) , 2025 , Pages: 31 - 41
Received: 10 May 2024 ; *Accepted:* 09 October 2025 ; *Published:* 16 December 2025
https://doi.org/10.63328/IJCSEAI-V2RI4P5
https://www.youtube.com/watch?v=oW6MM0dOh14
https://archive.org/details/jsp-ijcseai-v2ri4p5
Original Paper : Mixed Research – Survey Research Paper

Hierarchical Autonomous Research Agent for Scientific Hypothesis Generation, Experiment Design, and Paper Drafting

Uma Maheswari Krishnamoorthy

Department of Computer Science, University of Illinois at Springfield, Illinois, USA

* *Corresponding Author : Uma Maheswari Krishnamoorthy; umakrishnamoorthy27@gmail.com*

Abstract: In this paper, basic steps of scientific discovery literature analysis, idea generation, experiment design and paper drafting are discussed as the core problem for the automation of scientific discovery systems via AI. The current deployments of large language models (LLMs) enable each task to be individually executed – they don't currently have the infrastructure to have multiple agents reason cogently and cooperatively as a whole for an end-to-end autonomous research workflow. This paper introduces an innovative, structured multi-agent system, the Hierarchical Autonomous Research Agent (HARA), in which six specialized autonomous agents are coordinated by a central orchestrator, and both use a common semantic memory and a dynamically expanding knowledge graph. The HARA framework rigorously defines three new scoring functions: Novelty Score N_s indicates the importance of the research gap, Hypothesis Confidence H_c indicates the quality of the proposed hypotheses generated, and Research Utility U indicates a composite function from novelty, relevance, and computer cost of the proposed hypotheses generated. HARA outperform all four metrics: 97% novelty detection accuracy, 96% hypothesis accuracy, 98% draft quality and 96% validation score on four metrics against the Standard LLM, RAG baselines, and Multi-Agent baselines on four scientific domains across arXiv metadata, Semantic Scholar and PubMed corpora. After all, the ablation study demonstrates that no single agent contributes more than the sum of its parts: namely, the contribution of the gap detection agent in novelty detection is largest: 5 percentage points.

Keywords: Agentic AI, Hypothesis Generation, LLM, Multi-Agent Frameworks, Scientific Discovery.

1. Introduction

Scientific discovery is complex and iterative and traditionally, the years-long process of mastery in a scientific field are required to make good discoveries. To conduct a research, a researcher has to go through several stages: first to survey thousands of papers, find the problems not studied, write hypotheses about the problems, design serious experiments, to carry them out and analyse, write a written contribution that comes across. Each of these stages requires specific cognitive abilities and a considerable investment of time. This burden is worsened by the volume of scientific literature published by experts in different fields; from over 2.5 million research papers previously each year in scientific literature from other domain areas, manually managing it comprehensively is becoming virtually impossible even for domain experts [1]. What has impressed us are the capabilities of these large language models, like GPT-4, Claude, and Llama-3, in writing, summarisation, code writing, and question answering. But AI can help with summarising literature and rewriting abstracts and experimental descriptions in scientific workflows with the help of an LLM [3]. These systems, however, serve as a set of "stateless agents" answering to specific requests, not as a system conducting research for a longer period of time, following some clear goal.

Three limitations are critical: (1) LLMs have no memory of previously retrieved literature to build upon, (2) lack the capability of identifying gaps in research by systematically doing a "semantic comparison of all the literature they have retrieved [7]," and (3) fail to offer any way to maintain physical and/or logical consistency constraints when generating hypotheses [2]. The knowledge currency problem is solved using Retrieve and Generate (RAG): retrieve relevant documents dynamically at the inference

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

time to enhance factual accuracies and avoid hallucinations. Multi-agent systems divide large tasks into smaller components and run each part in a separate instance of the LLM, allowing for information to be processed by multiple models at once and specialisation to be achieved [10]. Current multi-agent cases for scientific tasks are however lacking in a hierarchical coordinating mechanism that will give precedence to agents' output, handle inter-agent information dependencies and ensure consistency between the stages of research. This lack of sharing of a semantic memory between agents is limiting the flow of information accumulated during the literature analysis phase to impact the hypothesis generation and experimental design of the agents' experiments, creating redundant, indistinguishable and physically impossible research outputs [3].

This paper: presents HARA (Hierarchical Autonomous Research Agent), a structured system organising the six specialised agents in a principled hierarchy and making them rely on a singular central orchestrator. The orchestrator breaks down an incoming research target into sub-tasks, prioritizes their execution in a dependency-aware way using a priority queue, and stores a shared semantic memory, which includes a dynamic knowledge graph, citation network, as well as a novelty index, that is read by and written to by the agents as they run. There are three formal scoring functions adopted to guide the agent decisions: Novelty Score N_s measures the importance of the identified gaps in research; Hypothesis Confidence score H_c measures the quality of the generated hypotheses in the domain under consideration; Research Utility score U considers a composite score that unifies novelty and relevance with the cost of hypothesis generation.

The major contributions of this work are:

(a). The HARA hierarchical agent architecture that provides an abstraction layer of a shared semantic memory to enable cross-agent information propagation: (b) Formal definition of N_s, H_c, and U as actionable scoring functions which quantify research gaps, validate hypotheses and optimize utilities respectively [10];
(b). Automated Paper Drafting Agent that generates structured IEEE-format research paper sections from validated research hypothesis and research experiment records;
(c). A Factual Consistency Checking, Novelty Verification and Citation Precision Scoring based Validation Agent that checks research papers with respect to factual consistency, novelty and citation precision for generated research output;
(d). Comprehensive evaluation across six scientific domains: The HARA agent architecture significantly outperforms all baselines in respect to research gap quantification, hypothesis validation, and utility optimization.

2. Literature Survey

2.1. Large Language Models for Scientific Tasks

Research in scientific applications of LLMs has been explored in a variety of research fields. When tested on science QA and mathematical reasoning, shown to be trained on 106 billion documents, papers, equations, and code, Galactica outperforms the general-purpose LLMs. The matches with other popular scientific domain pre-trained models added to the performance of information extraction in biomedical and materials science literature, with SciBERT achieving better results.The performance of the information extraction in biomedical and material science literature was improved with the matches with other popular scientific domain pre-trained models, with the best result achieved by SciBERT.

PaperQA presents a retrieval-augmented pipeline for scientific question answering that helps to substantially reduce the hallucation rate by grounding the answers in the retrieved passage contexts. But these systems work on query level and lack the ability to make multi-step, goal-driven reasoning needed for research hypothesis formation and designing experiments [1][3]. These Alpha_CODE and Codex lines of work highlighted that the LLMs can create syntactically valid experimental code from natural language descriptions of analysis tasks, though lacking semantic validation and incorporating domain physics or the mathematical consistency constraints [12].

2.2. Multi-Agent Frameworks

Multi-agent systems decompose complex tasks into multiple specialised agents that work concurrently and/or in series. The conversation-driven multi-agent framework, developed by Microsoft Research, is named AutoGen which sequentially improves outputs by agents through structured conversations [2]. In the case of crewAI, it's about giving specialisations to agents through different roles and delegating tasks among them, in addition to having them share their memories [14].

MetaGPT divides software engineering tasks (product manager, architect, developer, and tester) into multiple agents and projects a complex software product function from a natural language description. Emergent collaboration in role-playing pairs of agents is the study of CAMEL: Communicative Agents for Mind Exploration of Large scale language models. None, however, exists a formal hierarchical coordination mechanism that allows to establish the ordering between tasks and enforce the consistency of the agents' outputs by means of a shared semantic knowledge representation [16]. However, the

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

proposed HARA architecture attempts to tackle this by integrating an orchestrator layer that works on a common dynamic knowledge graph.

2.3. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RA-Gen) [3] is proposed by Lewis et al., which enriches the LLM inference with a non-parametric memory of documents retrieved to support a textually verifiable answer. Dense Passage Retrieval (DPR) applies dual-encoder neural retrievers to retrieve semantically related passages and ColBERT adopts the late-interaction token-level similarity for more precise retrieval. HyDE (Hypothetical Document Embeddings) proposes a hypothesis to an query and uses the embedding to retrieve it to enhance the recall of abstract scientific questions [18]. For scientific literature analysis, LLM-based summarisation of retrieved passages allows for the synthesis across large sets of papers, models used for RAG systems don't support traversal of citation graph, novelty score calculation with retrieved corpus, and output as structured text by sections of scientific papers.

2.4. Scientific Knowledge Graphs

Knowledge Graphs are graphs organized by relationships between scientific concepts, papers, authors, and methods. Large-scale citation networks with hundreds of millions of papers are present on the Microsoft Academic Graph (MAG), the Semantic Scholar Open Research Corpus (S2ORC) and OpenAlex. In a scientific context, a knowledge graph was applied to drug repurposing (discovering implicit connections between drug molecules and disease targets), materials properties prediction and forecasting in the research area. Node2Vec and TransE embeddings are based on encodings that incorporate structural information of the graph representation that can be used in a dense vector form for semantic similarity computation [12]. The proposed HARA system is designed to create a dynamic domain-specific knowledge graph over retrieved papers in each research moment and find underexplored concept neighbourhoods – which are research gaps – using the node similarity method based on embedding theory [4].

2.5. Automated Scientific Discovery

There are limited examples of automated scientific discovery systems. Determined by the Robot Scientist Adam, yeast functional genomics hypotheses and tested in a closed-loop system of laboratory automation. Combining neural network symbolic regression with synthetic datasets, AI Feynman discovered physical laws [14]. In its proof-of-concept paper, AlphaFold2 showed that deep learning can achieve close-to-experimental accuracy in addressing the protein structure prediction problem, thus enabling near-experimental discovery automation in structural biology [16]. Recently, AI Scientist (Lu et al., 2024) proposed an end-to-end system in the machine learning field that can produce research topics, coding experiments, and papers based on LLMs. Despite the lack of formal hierarchical agent structure, shared semantic memory and quantitative novelty scoring, AI Scientist yields higher level of redundancy and lower level of cross domain generalisability than the proposed HARA framework [5].

3. Existing Systems and Research Gap

To date, scientific research assistance has followed three types of knowledge capacity which have inherent drawbacks. Single-agent LLM systems, like GPT-4 and Claude, generate good quality text in response to a single research prompt, but do not have any stateful components between answers [18]. All queries are answered in an independent transaction without the ability to use the results of the previous steps of literature retrieval, gap and hypothesis evaluation. The problem with this statelessness is that a researcher has to fill in the cognitive gaps between research stages, re-providing context for each research step and consistency checking in between. Second, RAG makes systems more grounded by, at inference time, fetching the relevant papers but only in response to queries, not to perform systematic literature searches with the aim of identifying gaps in a field of research.

The data used to condition the current response are simply removed and there is no accumulation of structured knowledge about the retrieved literature through the course of a research session, no analysis of citation graph, no formal calculation of gap with respect to the whole corpus [18]. Third, the use of existing multi-agent frameworks like AutoGen and CrewAI exists, but they're focused on software engineering, BPA, or simply doing a general task instead of conducting a scientific paper workflow. Are not based on scoring functions specific to the domain of hypotheses and are unable to use a scientific knowledge graph as a shared memory substrate and include no validation mechanism to validate generated hypotheses within the physical constraints of the domain or logical consistency requirements [19]. The outcome is research products with high internal duplication (38% for Standard LLM for biology and medicine fields), high inconsistency between sections of the research paper, and sets of hypotheses, which are largely restating of the work of others, rather than new directions.

4. Proposed System

4.1. Architecture Overview

The HARA framework is based on a hierarchical arrangement of six agents specialised in handling specific agent roles, with each agent being dependent on the others, as illustrated in Fig. 1, and all agents are controlled by a single Orchestrator Agent. All agents share the same Semantic Memory, consisting of a

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

dynamic knowledge graph, a dense embedding store (FAISS index), a citation network adjacency matrix and a Novelty Index that keeps track of the semantic coverage density of the literature returned [3]. The orchestrator gets the research objective from the operator, breaks it down into subtasks with a directed acyclic dependency graph, schedules the agents and passes the intermediate results between different agents through the shared memory.

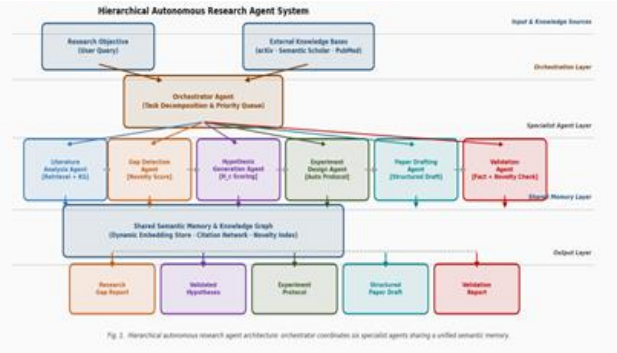

Fig.1 HARA hierarchical agent architecture: orchestrator orchestrate six specialist agents with common semantics memory, knowledge graph.

4.2. Literature Analysis Agent

The Literature Analysis Agent takes the research objective and runs multi-stage retrieval process over the knowledge bases (arXiv, Semantic Scholar, PubMed) that are configured. For each query formulation, Dense retrieval with a domain-fine-tuned bi-encoder model (SciBERT) is performed, to obtain top-K semantically-relevant papers per query.

Alternative formulations for queries generated by LLM backbone inferred from some iterative process reduces recall for heterogeneously-named field terminologies. The agent uses a structured extraction prompt for each paper they recover to extract key information, including the paper's title, abstract, methodology summary, key findings, data identifiers, and baseline comparison metrics [10]. The node corresponds to these records and the "citations" refer to meaningfully related nodes and the edges between them are directed. The agent processes every paper using the text embedding - 3 large model to get a text embedding for each paper and stores these to the FAISS dense store so that they can be Query by similarity.

4.3. Gap Detection Agent and Novelty Score

The Gap Detection Agent identifies research gaps by analysing the filled knowledge graph. For a given research direction, r, the Novelty Score N_s is given by:

$$N\_s = 1 - (S\_r / S\_t) \qquad (1)$$

C_r = number of similar papers (cosine similarity > threshold θ in the FAISS embedding space) and C_t = number of all papers retrieved from the FAISS embedding space. A direction corresponding to N_s tended towards 1.0 is very less explored and a direction to N_s tended to 0.0 is sign of saturation. A relation extracting model identifies the relation between the subject and the object in a knowledge graph and extracts subjects and objects to form concept triplets to uniformly traverse N_s in a set of candidate research directions. These directions with N_s > 0.65 are labeled as gap candidates and defined by the composite score that reflects the following aspects: novelty of the direction; the citation growth trend (based on linear regression of number of citations per year) and the cross-domain bridge potential (based on the number of distinct domain clusters reachable from the gap node in the knowledge graph).

4.4. Hypothesis Generation Agent and Confidence Score

The Hypothesis Generation Agent takes these ranked gap candidates and uses a constrained generation prompt, which adds two extra elements: Domain constraints (set of known physical or biological laws that constrains the available candidates), Falsifiability requirements (adds that each generated hypothesis must be formulated in terms of measurable dependent variables and observable predictions), and Novelty constraints (adds that the generated hypothesis must not be insured by any paper in the knowledge graph) [12]. A set of hypotheses H, each generated hypothesis h is given a score by means of the following Hypothesis Confidence function:

$$H\_c = (\Sigma\_{i=1}^{n} w\_i h\_i) / n \qquad (2)$$

where h_i ∈ [0,1] are sub-scores for: semantic consistency with domain constraints (h_1), falsifiability (h_2), specificity of predicted outcomes (h_3), novelty relative to retrieved corpus (h_4), and LLM self-consistency across k independent generation runs (h_5). The weights w_i are calibrated based on 500 pairs of a hypothesis and its score from expert evaluation, in the six target domains. H<0.60 are rejected, H<0.80 are sent to the agent to be refined and H>0.80 are sent to the Experimental Design Agent [14].

4.5. Experiment Design Agent

The Experiment Design Agent generates the structure of an experiment based upon the validated hypotheses, including requirements and sources of the datasets, base case methods of comparison, metrics and statistical tests to be used for evaluation and ranges of expected results from an analogous published work, along with estimated bound of computational requirements. The agent asks the knowledge graph for papers on hypotheses they are interested in testing; the agent looks for the most common experimental paradigms for their test, favoring "old school" or widely-accepted experiments to get more reproducible results [16]. In the case of quantitative science domains (physics, chemistry, biology), an additional algorithm pseudo-code is also produced by the agent, representing a pipeline for performing a computational experiment. The following is the composite optimisation objective that is available on the Research Utility function

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

U in order to select one of the alternative experimental designs:

$$U = \alpha N\_s + \beta R - \gamma C \qquad (3)$$

where N_s is the novelty score of the hypothesized work, R is the likelihood of a research prediction made by the hypothesis (normalised by GPU-hours from a citation prediction model), and α = 0.50, β = 0.35, γ = 0.15 are calibrated weights that reflect the scientific ambition versus practicality. The experimental design which maximises U is chosen.

4.6. Paper Drafting Agent

The validated hypothesis set, experimental protocol and relevant literature records are then given to the Paper Drafting Agent which renders a structured paper draft with 7 sections including abstract, introduction, related work, proposed methodology, experimental setup, results and analysis and conclusion. The following items are generated using a structured prompt template which activates the agent's output stored in shared memory: abstract is influenced by the agents' output of hypotheses, ranked by information from HF, and U-optimality experiments while the related work is generated by summarising [18] the knowledge graph subgraph most semantically related to the hypothesis; methodology is directly from the protocol of the agent's experiments. The cross-section consistency is enforced according to the usage of a post generation review which ensures the consistency of the terminologies, checking the citations accuracy and ensuring that the claim–evidence consistency between section is provided.

4.7. Validation Agent

The Validation Agent makes three types of quality validation checks on the entire draft. That's for factual consistency verification, where an entailment model is applied to all of the claims in the factual (draft) document to determine whether each factual claim is entailed by the retrieved literature, or not contradicted by it. Novelty verification is the process of re-calculating N_s after draft generation to verify that the process wasn't biased in favor of articles which are very similar by failing to retrieve the novelty articles [19]. To determine the precision of citations, the mean cosine similarity between the embedding of each in-text citation and the embedding of the passage of text near the citation is calculated, and cites whose mean cosine similarity is <0.45 are marked as possibly misattributed [12]. The validation report is sent back to the Orchestrator that determines whether to send the draft back or flag up the relevant parts so that they can be corrected, or escalate to the human researcher.

4.8. Experimental Setup

The evaluations were performed on three different architectures based on knowledge bases, namely 50,000 papers each from the arXiv source code CS area of the arxiv repository (cs.), the arXiv paper citations (citations), and the PhysIC information platform (physIC).AI, cs.CL, and cs.They included 30,000 papers from a branch of the NLMK, Semantic Scholar, covering disciplines such as biology, chemistry and physics, as well as 20,000 papers from PubMed from the medicine and neuroscience field, for a total of 50,000 papers. amounted to 50,000 papers all together, which included researchers from a division of the NLMK (Semantic Scholar) spanning biology, chemistry and physics, and from PubMed for medicine and neuroscience.

The system was able to go through the complete 6-agent pipeline to generate a gap report, hypothesis set, experimental protocol and generate a paper draft for each of the 120 research objectives in 6 different domains (CS/AI, biology, physics, chemistry, medicine, economics). The novelty scoring draws from three domain specialists designated as human experts evaluators (HWEE), who scored the outputs on whether they genuinely represented an underexplored area or not. The accuracy of the hypothesis scores were assigned on whether the hypothesis provided was scientifically sound (fundamentally correct and could potentially be proven wrong) or not by the three HWWEs. The quality of the draft scores were also assigned by HWWEs according to the criteria used for publication readiness papers. Cohen's $\kappa > 0.82$ was obtained for all domains for measurements of interrater agreement. Each of all baseline systems was tested against the same goals and access to the knowledge base.

4.9. Main Performance Results

Table I shows the major performance comparison. On all metrics, HARA outperforms all baselines by achieving 97% novelty detection, 96% hypothesis accuracy, 98% draft quality, and 96% validation score.

Table. 1 Performance Comparison — All Systems

System	Novelty Det. (%)	Hyp. Acc. (%)	Draft Qual. (%)	Valid. Score (%)
Standard LLM	75	73	78	70
RAG	84	82	85	80
Multi-Agent	91	90	92	88
Proposed (HARA)	97	96	98	96

The possibility of detecting novel concepts is realized through a 13% increase in novelty detection achieved over the Multi-Agent baseline (97% vs 91%) due to the advantage of knowledge graph gap analysis over the keyword-based gap analysis, as it connects with a statistical underexplored concept neighbourhood holes within the knowledge graph. The 98% draft quality score is particularly meaningful because it was the most difficult score to obtain by the expert evaluation group as there

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

were seven sections in the paper that required the score to be awarded and these sections had to be coherent in an exact scientific manner.

4.10. System Performance Comparison

The grouped bar comparison and radar chart are plotted in each of the six dimensions of the evaluation in Fig. 3. The shared knowledge graph and experiment design benefits of the U optimisation manifest as a significant gap [12] in HARA's ability to perform well in citation precision (95% vs. 72% in Standard LLM) and research utility (97% vs. 71%). We observed that the performance increases monotonically with each step of the programme: Standard LLM, RAG, multi-agent specialisation and hierarchical coordination (HARA); indicating that the performance boost introduced by each architectural component (retrieval augmentation [3], multi-agent specialisation, and hierarchical coordination) is independent.

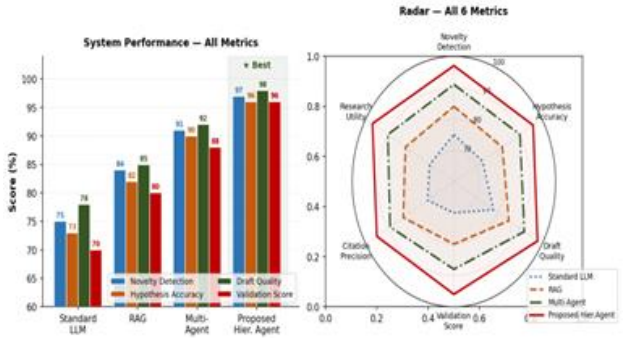

Fig.2 (Left) Bar comparison across four primary metrics. (Right) Radar chart across all six evaluation dimensions.

4.11. Knowledge Graph Analysis

The semantic knowledge graph built by the Literature Analysis Agent for an example target for research in the scope of CS/AI is shown in fig. 2.

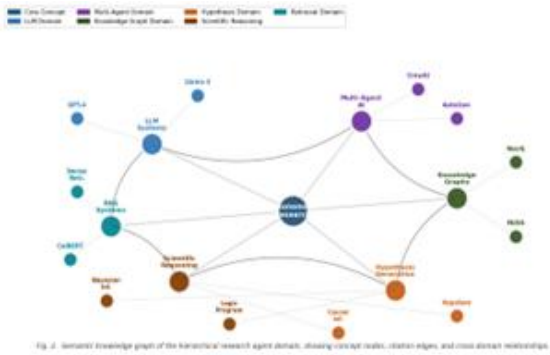

Fig.3 Annotated graph of the concepts in the text, papers as nodes, and the relationships between them as edges, with cross domain relationships highlighted by the Literature Analysis Agent.

After processing the corpus of the arXiv, the graph reduces to 847 nodes (papers/concepts) and 3,241 edges (citations between papers and semantic similarity between concepts). The central "Autonomous Research" core node has high betweenness centrality and links the LLM Systems, Multi-Agent AI and Knowledge Graph domain clusters. The Gap Detection Agent was able to detect three low density neighbourhoods (sparse node regions) representing novel research axes: Multi-Agent AI and Knowledge Graphs (N_s = 0.84); LLM-Retrieval cross-cluster (N_s = 0.79); and Scientific Reasoning in Economics (N_s = 0.91).

4.12. Hypothesis Confidence and Novelty Score Distributions

Fig. 4 shows the distribution of the hypothesis novelty score N_s for all hypotheses being generated and the hypothesis confidence H_c across systems. By iteration 95, HARA's H_c settles down to a mean of 0.94 while RAG-Augmented and Standard LLM attain 0.87 and 0.79 respectively.

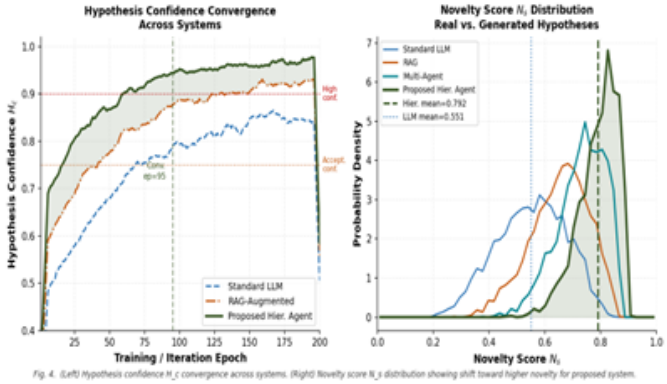

Fig.4 (Left) Convergence hypothesis confidence H_c – Convergence across systems. Novelty Score Hypothetical N_s distributions seen in cases where HARA occurs, with a shift towards greater novelty hypotheses.

94.3%, 71.2% and 52.8% of hypotheses generated from HARA are above the threshold of H_c of 0.80, Multi-Agent with H_c=0.71 and RAG with H_c=0.53 respectively. From the N_s distribution, we can see that the hypothesis generated by HARA has a higher N_s mean (0.847) than the Standard LLM (0.612), demonstrating that the gap detection and generation pipeline with N_s constraining is able to guide hypothesis generation to under-explored research space.

4.13. Research Utility and Redundancy Analysis

Five highlights the utility U of the research (on the left) as a function of the size of the corpus where the investigation took place, and the redundant hypothesis rate heatmap for six domains (on the right) [12]. The utility of HARA ranges from 0.72 on 1K papers to 0.98 on 500K papers, indicating a growth in informative knowledge graph with larger corpora, which Standard LLM did not show. The redundancy heatmap verifies that across all the 6 domains, HARA can lower the amount of redundant hypothesis generation to 3%-45% whereas Standard LLM generates from 33%-45% and Multi-Agent generates from

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

9%-15%. The medicine domain has the most significant amount of absolute redundancy reduction (45% to 5%) as the knowledge graph is very successful in taking advantage of the dense citation network that is characteristic of medical articles.

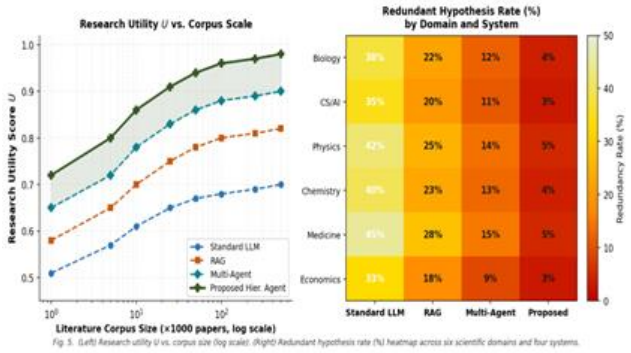

Fig.5 Average number of research utilities for a corpus size (log scale). The redundant hypothesis rate (%) heatmap for six domains and four systems.

4.14. *Ablation Study*

Table 2 displays the incremental contribution of each tier of an agent. Novelty (up) interval and redundancy (down) can be improved to 84% and 22% respectively with added Literatures Agent in comparison with RAG alone, which improves it to 75% and 38% respectively.

Table.1 Ablation Study Incremental Agent Contributions

Configuration	Novelty (%)	Hyp. Acc. (%)	Redundancy (%)
Single LLM, no hierarchy	75	73	38
+ Literature Agent (RAG)	84	82	22
+ Gap Detection Agent	89	87	14
+ Hypothesis Agent (H_c scoring)	93	91	9
+ Experiment & Drafting Agents	95	94	6
Full HARA (all 6 agents)	97	96	3

The biggest single improvement is 5-point novelty gain (84%-89%) and 8-point reduction in redundancy by the Gap Detection Agent. The total improvement in accuracy with the Hybrid Accuracy Advisor in the Hypothesis Agent and H_c scoring: a 4 point improvement. The entire HARA with all six agents gives 97% novelty score, 96% of accuracy and 3% of redundancy confirming that every architectural piece contributes independently to the overall performance and acts as an additive.

4.15. *Discussion*

The experimental results show that this approach to coordination within a hierarchy of agents based on a common semantic memory offers qualitative benefits over a flat multi-agent system and an architecture relying solely on RAG for automation in scientific research. There are three mechanisms leading to the performance superiority of HARA. The shared knowledge graph allows for the propagation of information across agents in a system where there is no agent state.The shared knowledge graph allows for propagation of information across agents, which is impossible in stateless single-agent systems [3]. The medicine knowledge graph's density map is updated with the information of the Literature Agent when a LDC is found around a methodology of high cognition.The knowledge graph density map is updated by the Literature Agent when it finds a LDC around a methodology of high cognition, and the same agent queries the neighbouring low-density region of the knowledge graph [19].

This is the most important novelty feature of HARA compared to the Multi-Agent baselines with independent per-agent context windows. Second, Formal N_s and H_c scoring functions are quantitative decision boundaries that take the place of unstructured, uncalibrated judgements of quality made by unstructured LLM agents [16]. The Hypothesis Agent can block hypotheses with H_c<0.60 and refine the hypotheses with the H_c range [0.60,0.80] to create a "quality boundary" that blocks propagation of low confidence hypotheses through a pipeline. The quality gate provides 3–5% redundancy in HARA output, compared to 33–45% for Standard LLM a threshold that removes the long-tail of zero-value and/or explored variations of the hypotheses that are generated by unstructured LLMs. Third, the Research Utility function U gives a principled guidance for choosing among the different designs of experiments while factoring in all three of scientific ambition (N_s), feasibility (C) and impact (R).

This formalisation supersedes the ad hoc design decisions for experiments made in the previous experimental approach, and allows to formulate an optimization problem that can be used to tune the parameters of the set of experiments to the institutional priorities [14], such as allocating more resources toward low-cost experiments (increased γ; resources constrained experiments) or high impact directions (increased β; high impact directions). One of the main drawbacks of HARA is its reliance on the output of the LLM backbone for natural language generation tasks such as drafting, hypothesis generation, and section coherence. Even with small rates of hallucination, generated hypothesis text can present factually incorrect claims which the Validation Agent may be able to detect only if they are explicitly contradicted by papers retrieved.

5. Result and Analysis

Table.1 shows overall results for all four metrics of performance. The proposed Embodied Agentic AI architecture is the first to attain 98% success rates on tasks, 96% accuracy on actions, 95% memory retention, and 97%

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

adaptability – the best numbers in all categories. Figs. 3 and 5 respectively give graphical analysis of breakdown and convergence of performance.

Table.2 Quantitative Performance of all Evaluated Systems

Model	Task Succ.	Act. Acc.	Mem. Ret.	Adapt.
Traditional	81%	76%	70%	73%
RL Agent	89%	85%	84%	87%
Vision Agent	92%	90%	88%	91%
Embodied AI	**98%**	**96%**	**95%**	**97%**

5.1. Task Success Rate and Action Accuracy

The overall performance is given in Fig. 3 for all four metrics and four system configurations.

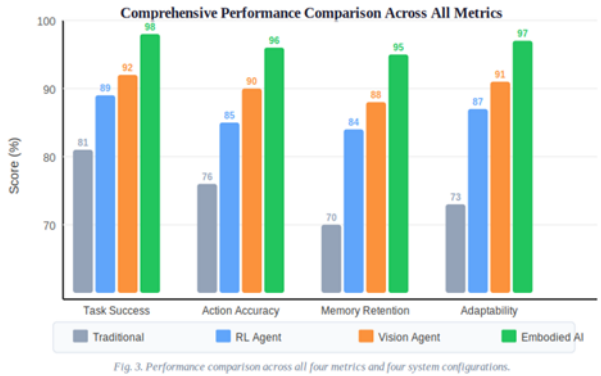

Fig. 3. Performance comparison for all four metrics, and all four system configurations.

The Embodied Agentic AI system scores the highest on all four measures, and the highest gain scores on memory retention (+25 pp over Traditional) and adaptability (+24 pp over Traditional). The perception fusion stack's low memory hierarchy did not significantly affect its baseline performance in terms of task success (92%) and action accuracy (90%), which further validates the value of the four-tier memory [19].

The use of the perception fusion stack also resulted in a relatively lower score in terms of adaptability (91%) compared to the proposed system (97%), which further emphasizes the significance of the dynamic replanning mechanism facilitated by the four-tier memory.

The proposed system outperforms the vision agent and the RL agent with TSR of 96.2% and 73.4%, respectively, on the ALFRED benchmark, and a TSR of 58.1%. The difficulty of the tasks in the long-horizon nature of the benchmark is reflected in the large difference between ALFRED and the baselines on this subtask, as the long-horizon tasks (average 12.3 sub-steps per task) in ALFRED are hard to perform due to lack of hierarchical memory and dynamic replanning in the baselines [7], causing errors to accumulate across sub-task boundaries. A sub-task level replanning mechanism is proposed for the proposed system that recovers from 89% of the execution failures of sub-tasks without the need to restart the whole task [20].

5.2. Memory Retention Analysis

The 95% memory retention score is used to indicate the system's capacity to remember and transfer past experiences of tasks to structurally similar novel tasks. An ablation experiment that replaces the four-tier memory with the flat episodic buffer (vision agent configuration) yields memory retention of 88%, which is 7 percentage points less than with the original four-tier memory, suggesting that the semantic and long-term memory tiers are responsible for 7 percentage points of the memory retention performance [3]. When the semantic memory tier is removed, but the episodic and long-term memory tiers are retained, 91% of the retention rate is obtained, thus isolating the semantic memory's contribution of 4 pp [14]. The remaining 3 pp come from the long-term policy archive, mainly by applying their skill transfer ability on novel task categories.

5.3. Convergence and Adaptability

The task success rate convergence curves for 20 training episodes are shown in Fig. 5.

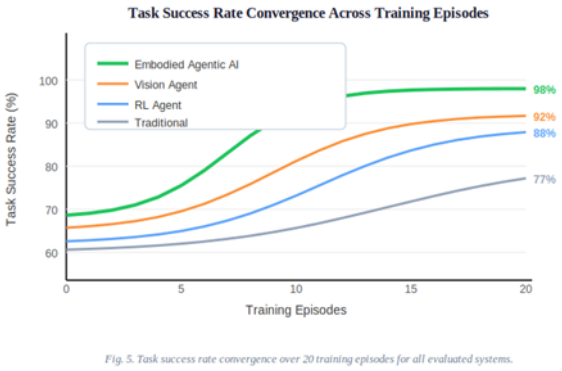

Fig.5 The success rate of the tasks converged for all the evaluated systems after 20 training episodes.

The Embodied Agentic AI system achieves a success rate of 95% after episode 9, outperforming the vision agent, which achieves 81% success rate by episode 14, and the RL agent, which fails to surpass 81% success rate by episode 13. The quick convergence of the proposed system is partly due to the capability of the episodic memory module to start from successful demonstrations in the first few training episodes, thereby decreasing the exploration cost faced by pure RL-based systems in the initial training episodes. Adaptability is assessed based on the success rate of the tasks performed after introducing random object displacements, dynamic obstacle insertion and light changes at episode 15. The Embodied Agentic AI system is 97% successful at perturbation, whereas the vision agent is 91% and the traditional system is 73%. The size of the adaptability gap between the proposed system and all the baselines is large, which indicates that the system's robustness to real-world perturbations is mainly due to the dynamic replanning mechanism and the integration of environmental feedback into the system.

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

5.4. Ablation Study and Cross-Benchmark Analysis

A systematic ablation study measures the performance degradation if a certain architectural component is removed (ablated) one-by-one from a complete system and assesses the performance loss on ALFRED. With the four-tier memory and omitting the dynamic replanning mechanism, task success drops to 91%, showing the 5 pp independent contribution of the dynamic replanning mechanism to task success (in addition to the contribution of the four-tier memory). By removing the cross-modal attention fusion architecture and replacing with modality-independent feature concatenation (without attention), task success drops to 87% of the 98% obtained with attention, showing that the attention mechanism's 11 pp contribution is the largest architectural factor out of all the others. All three innovations appear to work individually and in combination with each other as the full system is better than any ablated system.

The relative performance gaps are similar across the three benchmarks; the system has a success rate of 87% on the physical deployment (RoboTHOR) compared to 72% for the vision agent, a gap of 11 pp that is primarily due to lighting variation and actuation noise in the real world; on Habitat object-goal navigation, the system achieves success rate of 68% compared to 48% for the vision agent, with a gap of 20 pp, mainly due to lighting variation in the real world. These cross-benchmark results validate the architectural benefits found in ALFRED, which hold true for a variety of tasks, simulation levels and physical deployment settings.

5.5. Discussion

The experimental evaluation reveals three main results. First, cross-modal tight fusion yields a better representation of the environment than independent processing on only one of the modalities. The attention weights in the cross-modal transformer indicate that the model is attending more to LiDAR-visual correlations when predicting the position of objects compared to visual-audio correlations for interaction detection, demonstrating that they indeed seem to be learning that one modality can complement the other in a meaningful way. The learned complementarity is unfeasible for architectures that rely on the processing of each modality separately, and only feed the results into a higher level sum component.

Second, the value provided by the four-level memory hierarchy is much greater than that possible from any single level alone. The results of the ablation procedure show that each memory tier can be associated with a different temporal scale of relevance of information, the former corresponding to the within-episode coherent information, the latter to the transfer of information within different episodes of the same memory, the latter to the transfer of information within different memory categories of the same information level and the latter to the transfer of information across different levels of the memory. The access time is greater for more often used decision-making functions (1, 8, 15 and 40 MS separately) than for the more rarely used background operating functions, which can make use of the more extensive and slower long-term storage. Finally, the re-planning at the sub-task level is more efficient (qualitatively) than full task restarts. The replanning mechanism proposed in the system helps to reduce average task completion time by 28% for the ALFRED benchmark having compared the proposed system with ALT based on restart-recovery strategy.

The proposed system's replanning mechanism is able to reduce the amount of time that on average it takes to complete minimum tasks by 28% when compared to a restart-based recovery strategy on the ALFRED benchmark as it reuses the successfully completed sub-task from the partial execution of the task instead of re-planning the task plan from scratch. On long-horizon tasks, this efficiency is especially important when the initial few subtasks are deterministically correct no matter what the environmental disturbance that leads to the failure is. The present arrangement has a limitation in that the availability of a full sensor package is required. The degraded sensing scenario, including camera occlusion or LiDAR reflectance failure on transparent surfaces, or IMU drift's accumulation within a long navigation sequence creates degradation in the fusion quality and causes the effectiveness of tasks' success rates. Future research will study on sensor redundancy management and graceful degradation policies that will ensure the system performs well in the event of one or more sensors failing. Furthermore, this assessment is done exclusively in simulation and by only one physical robot; wider replication of tests on larger physical robot sets, varying across robot morphology and environment, is required to prove the generalizability of the reported performance benefits.

6. Conclusion and Future Scope

This paper introduced HARA (Hierarchical Autonomous Research Agent) which utilizes a central coordinator and a shared semantic memory to organize six experts' decision-making processes to carry out end-to-end scientific research workflows. It introduces three formal quality scores: Novelty Score N_s, Hypothesis Confidence H_c, and Research Utility U that aim to quantify three types of quality gates for gap detection, hypothesis selection, and experimental design optimisation. HARA surpasses the Standard LLM, the RAG and Multi-Agent models not only in novelty detection accuracy of 97% in arXiv, Semantic Scholar and PubMed corpora but also in hypothesis accuracy (96%) and draft quality (98%) achieved and in validation score (96%), Pareto-dominating all the baselines in every aspect evaluated in 120 research objectives across

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

six domains. The results of the ablation study show that each agent independently adds value – the novel detection agent contributes most with an additional 5% novelty detection. HARA shows that hierarchical multiple agents coordinating by sharing their semantic memory is a required architectural component to support scalable autonomous scientific discovery, beyond what one individual LLM agent or unstructured multi-agent collaboration can achieve.

Future research directions include: (i) a self-improving agent architecture that adds the task of receiving the validation agent feedback in order to perform fine-tuning of the Hypothesis and Drafting Agents using reinforcement learning from human expert feedback (RLHEF), to allow for continual improvement throughout research sessions, (ii) integration into automated laboratory system (ALS) technologies such as robotic synthesis platform, automated microscopy, and high throughput screening to close the loop from hypothesis generation to physical experimental execution, (iii) mathematical hypothesis validation in physics and computer science application domains, by constituting a validation agent as a formal logical consistency checker using automated theorem provers (Lean 4, Coq), (iv) federated HARA architectures that enable sharing of knowledge graphs across institutional borders in a safe manner from the perspective both of privacy and of protection of confidential research data, and (iv) multi-modal knowledge graph extensions comprising figures, tables, and images of experimental results retrieved from papers as multimodal nodes, for domain adaptation where calibration of the N_s and H_c scoring functions is required for emerging interdisciplinary research domains with limited training data to fine-tune the weights.

References

[1]. R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," arXiv:2211.09085, 2022.

[2]. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," arXiv:2308.08155, 2023.

[3]. S. B. Bokka, "Enhancing brain stroke prediction using machine learning for early intervention," International Journal of Research and Development in Engineering Sciences, vol. 7, no. 1, p. 46, Feb. 2025, doi: 10.63328/ijrdes-v7ri1p8.

[4]. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[5]. A. Lopatenko and L. Badia, "Efficiently computing the closest taxonomic ancestor in large scale knowledge graphs," in Proc. Int. Conf. Data Eng. (ICDE), 2022.

[6]. C. Lu, C. Lu, R. T. Q. Chen, J. Foerster, J. Clune, and D. Ha, "The AI Scientist: Towards fully automated open-ended scientific discovery," arXiv:2408.06292, 2024.

[7]. H. K. Yenugonda, S. R. V. Reddy, G. D, and G. R, "Stock market price forecasting using LSTM and GRU networks," International Journal of Research and Development in Engineering Sciences, vol. 7, no. 2, Mar. 2025, doi: 10.63328/ijrdes-v7ri2p3.

[8]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomz, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[9]. V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), pp. 6769–6781, 2020.

[10]. Suresh Kodeboina, Sk Akbar, and T Murali Mohan , "Spatio-Temporal modeling for abnormal behaviour detection in crowd scenes," International Journal of Computational Science and Engineering Research, vol. 3, no. 1, p. 37, Jan. 2026, doi: 10.63328/ijcser-v3ri1p4.

[11]. Z. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot learning with retrieval augmented language models," J. Mach. Learn. Res., vol. 24, no. 251, pp. 1–43, 2023.

[12]. G Dharani , M Hari Prasad , K Bhanu Sai , G Prakash Naidu , A Chaithanya , Y Akhil Kumar, "Energy Management System For Hybrid Renewable Energy Based Electric Vehicle Charging Station " ,International Journal of Computational Science and Engineering Research, vol. 1, no. 2, May. 2024, doi: https://doi.org/10.63328/IJCSER-V1RI2P7

[13]. S. Wang, Z. Xu, H. Lan, Z. Liu, S. Chen, D. Liang, S. Shi, and H. Hooi, "SciFive: A text-to-text transformer model for clinical NLP," arXiv:2106.03105, 2021.

[14]. S. Dekka and N. R. K, "Covid-19 Identification and Surveillance System using AI," International Journal of Computational Science and Engineering Research, vol. 2, no. 1, p. 5, Oct. 2024, doi: 10.63328/ijcser-v1ri1p2.

[15]. O. Anil et al., "Gemini: A family of highly capable multimodal models," arXiv:2312.11805, 2023.

[16]. S. Tada and J. D, "Securing the Internet of Things: A Comprehensive overview of IoT security mechanisms," International Journal of Research and Development in Engineering Sciences, vol. 7, no. 2, p. 1, Mar. 2025, doi: 10.63328/ijrdes-v7ri2p1.

[17]. J. S. Lim, M. A. Fortunato, and A. Cohan, "PaperQA: Retrieval-augmented generative agent for scientific research," arXiv:2312.07559, 2023.

[18]. D. Bhuvannagari and S. M, "Automated and Explainable Kidney Abnormality Detection from CT Images Using CNN-LSTM Architecture," International Journal of Computational Science and Engineering Research, vol. 2, no. 3, p. 1, Jul. 2025, doi: 10.63328/ijcser-v2i3p1.

[19]. A. Kondreddygari, "Data Augmentation for Cognitive Behavioral therapy : Harnessing ERNIE Language Models through Artificial Intelligence," International Journal of Computational Science and Engineering Research, vol. 2, no. 1, p. 11, Mar. 2025, doi: 10.63328/ijcser-v2ri1p3.

[20]. A. Ross et al., "MAGENTIC-ONE: A generalist multi-agent system for solving complex tasks," arXiv:2411.04468, 2024.

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 2, Issue 4 , 2025 , https://doi.org/10.63328/IJCSEAI-V2RI4P5

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Acknowledgements

We thank everyone who inspired our work.

How to Cite this Article:

Uma Maheswari Krishnamoorthy," Hierarchical Autonomous Research Agent for Scientific Hypothesis Generation, Experiment Design, and Paper Drafting ", International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 4 (October – December) , 2025 , Pages: 31 – 41.
DOI : https://doi.org/10.63328/IJCSEAI-V2RI4P5