BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Lightweight Multi-Agent Verification Framework for Trustworthy RAG to Reduce LLM Hallucinations

Shraddha Dattatray Amup

Department of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, Maharashtra, India
Shraddha.amup@dypic.in

Prof. Pooja Dehankar

Department of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, Maharashtra, India
poojadehankar@dypic.in

Dr. Vishal Puranik

Department of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, Maharashtra, India
vishalpura@gmail.com

Abstract

Although Large Language Models (LLMs) are mostly effective, they sometimes produce outputs that are not necessarily accurate, and Retrieval-Augmented Generation (RAG) offers a solution to the problem by getting the model to read some materials before producing the answer, but RAG may not be sufficient to resolve the issue completely. The retrieved sources may be anything from irreverent and contradictory to obsolete and unreliable. As a part of a standard RAG system, the verity of the retrieved information is not checked before giving the input to the language model. The present paper presents the results of ten studies dealing with multi-agent and multi-level verification of data, facts verification through hallucination detection, the process of responding to the debate, agent coordination, knowledge-graph and hypergraph-based retrieval, and the use of RAG-based reliable systems. The major issues include high costs of implementing the most effective method and the fact that in many cases verification is done post factum. To solve this problem, we proposed a simple model where only a few small agents check the relevance, reliability, and up-to-dateness of the quoted statements.The delivery of the passages occurs through a one-time single call, while there is no reply from the model in case all passages failed to pass the test. The project is more theoretical in nature, as we provide the description of the architecture, scoring model, algorithm, and evaluation plan, leaving the practical part of the project for the future.

Keywords: Retrieval-Augmented Generation, Hallucination, Multi-Agent Systems, Evidence Verification, Trustworthy AI, Large Language Models

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

I. INTRODUCTION

Tools such as ChatGPT, Gemini and Claude respond to inquiries, summarise texts and write programmes in a natural manner. They have a disadvantage known as the phenomenon of hallucination where the tools can provide information that is incorrect but phrased confidently. In the area of medicine, education, finance and law, even one misleading piece of information can lead to a negative outcome.

RAG tackles this challenge by allowing the model to consult with the retrieved documents prior to producing an answer. The shortcoming of this solution is that even if the relevant documents are present among the retrieved ones, the faulty output can still occur because the irrelevant or contradictory documents can be produced by the retriever. [1].

Problem. The current RAG approaches fail to apply any verification mechanisms prior to producing the answer. This means that while the already existing verification methods can guarantee the accuracy of generated information, they are inefficient, expensive and work only after the answer has been produced. In this context, it is important to create a cheap mechanism that can be applied before generation takes place.

Objectives. The objectives of this work are (i) to study hallucination and limitations of RAG, (ii) to analyze the most recent multi-agent verification studies, (iii) to analyze existing research work done in this field, (iv) to propose a light multi-agent verification framework, and (v) to specify how this framework should be evaluated.

Research Questions. RQ1. How can retrieved evidence be validated before providing it to a large language model? RQ2. Are small and simple agents capable of making an effective decision regarding relevance, consistency, freshness, and source reliability of the information? RQ3. Which metrics should show the effect of hallucination, and whether the answers are accurate and effective? RQ4. What limitations should be addressed before implementing the framework?

Scope. The paper reviews the literature and represents the conceptual framework for this study. No model was trained and no experimental results were reported.

II. BACKGROUND

Hallucination. An LLM comes up with a sequence of words by predicting the next possible token. As facts are not checked, a sentence can be grammatical with no evidence to support it. Hallucinations can go undetected, as they sound just as convincing as the right answers would.

RAG. A RAG system takes the user's query and finds the k passages from the document collection. The LLM uses the user's question and the selected text together to generate the answer. The knowledge can be easily updated just by changing the document collection without re-training the model.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

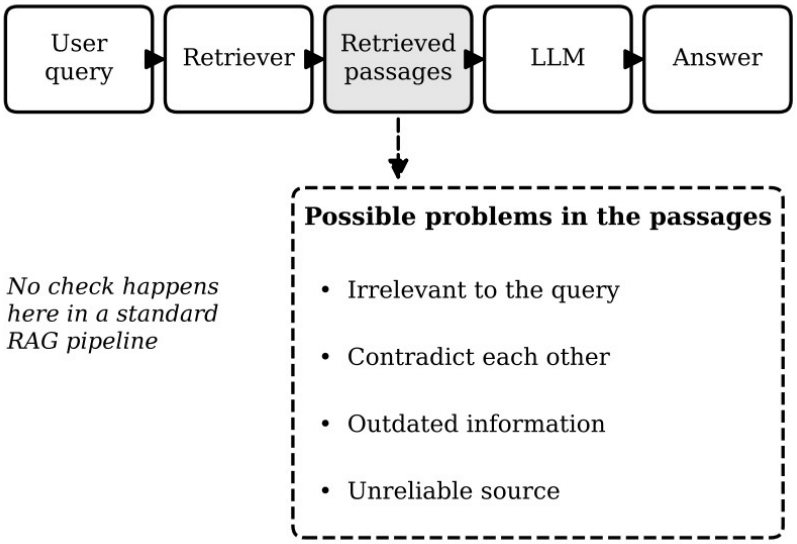


Fig. 1. A standard RAG pipeline and the problems that can occur in the retrieved passages.

Where RAG fails. There are four common cases. The passage might be irrelevant, might be contradictory with other passages, might be old, or might come from an unreliable source. Studies on trustworthy RAG consider each of the aspects of reliability separately, rather than treating it as a single property. [1].

Ways of minimizing hallucinations. Researchers have tried various solutions such as better prompts, training with reliable data, and grounding the model in external knowledge, as well as detecting false claims after the response has been generated. The approaches mentioned are viable, and each one has its cost. Training requires reliable training data and effort, while post-hoc detections are made only after mistakes occurred. Pre-emptive checking of evidence has been less studied.

Multi-agent systems. In this case, the tasks are performed by multiple agents, each of which is responsible for its own task. [2], [3]. Having different roles increases the reliability of the system but creates certain costs.[4].

III. LITERATURE REVIEW

We reviewed ten papers released during the years 2024 and 2026. The papers were found in the ACL Anthology, AAAI, NeurIPS, ScienceDirect, MDPI, Nature Communications and arXiv through keywords including hallucination, RAG, multi-agent LLM, knowledge graph, evidence verification and fact-checking methodology. We selected the papers that presented some clear methodology. The outline of the papers is shown in Table I, while the results are presented in different themes in Fig. 2.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

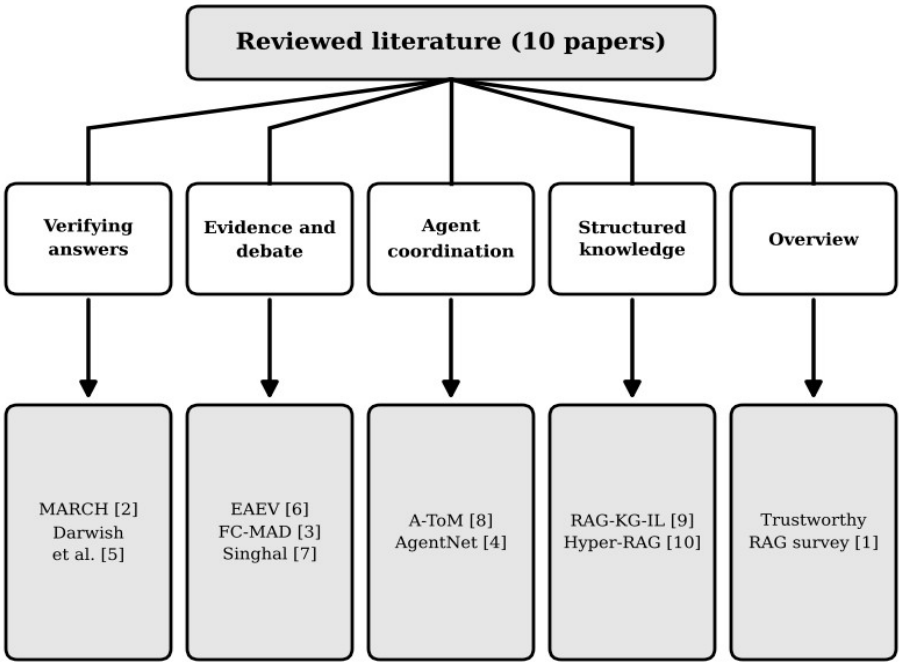


Fig. 2. Grouping of the ten reviewed papers into five themes.

Verification of generated answers. MARCH [2] employs three agents. Firstly, there is a Solver who generates the final answer, the Proposer who separates this solution into simple braced claims that can be validated against the evidence, and the Checker who verifies if each claim is true or false based on the evidence without having access to the original answer. This absence of information is intentional so as not to allow the Checker to merely confirm the previous error. All three components are taught via multi-agent RL, and it has been shown that an 8 billion-parameter MARCH model is comparable to several strong proprietary models. There is a downside to it- the effort in training. Darwish et al. [5] uses a combination of LLM and custom rules plus a verification loop in order to achieve stable results, although their rules are dedicated to one application only.

The alignment of evidence and ongoing debates. EAEV [6] identifies hallucinated entities based on how accurately the evidence corresponds to them in three ways and has incorporated a counterfactual stability test. It is useful in several RAG benchmarks but only establishes the presence of errors. On the other hand, FC-MAD [3] does not need any training and features many LLMs debating the statement, while a summary is concise so that a judge can come up with a decision easily. It passes a range of tests but increases latency in the process. However, Singhal et. al. [7] have developed a solution for producing the top three documents based on LLM-produced Q/A in order to properly classify claims and have reached a score of 0.33 on the Averitec dataset, which is a 22 percent absolute lift over a statement.

Agent coordination. Agent coordination. While these studies do not really address hallucination issue, they still provide insights related to the construction of multisource agent verification systems. The work A-ToM [8] demonstrates that coordination level may deteriorate if there are agents who reason at different levels about their partners, and this paper also attempts to define the reasoning level of the agents. The system

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

AgentNet [4] removes the role of a central controller and allows agents to coordinate their work in a dynamic graph with repository-based memory, thus exceeding the efficacy level of single-agent systems and systems with a central controller. The results obtained in these studies speak about the importance of coordination quality that increases with the growing number of agents.

Knowledge that is systematically arranged and retrieved. The discussion of the two articles highlights the efficiency of retrieving with the help of structured knowledge by also reducing the occurrence of hallucination. RAG-KG-IL [9] is a multi-agent hybrid system that includes RAG and knowledge graph combined with incremental learning, so the knowledge base does not need to be retrained in the future. It is reported about 73% less hallucination than GPT-4o in a healthcare case. Hyper-RAG [10] retains relationships among different entities in a hypergraph rather than in pairs. This technology enables the accuracy to be improved by 12.3% in average compared to using direct LLM and remains stable when the questions get more complex. The high price of this technology is the offline time needed for the knowledge base to be built through extraction, and RAG-KG-IL was only tested on a single healthcare case.

Critical analysis. There are three points that arise from a careful examination of the papers. First, verification works since methods verifying claims and/or evidence provide greater accuracy in terms of factuality than methods that do not use verification techniques. Secondly, the main benefits of verification methods diminish when their negatives have been taken into account. For example, MARCH requires training powered by reinforcement learning [2], FC-MAD uses multiple rounds of discussion between LLMs [3], and coordination studies affirm that the more agents are involved in multiple studies, the more burden will be on them [4], [8], and structured-knowledge methods necessitate constant building and updating knowledge bases [9], [10]. Thirdly, the majority of the methods are designed to verify the answer and not the evidence the answer is based on. For instance, EAEV uses a comparison of the entities in the generated answer and in the evidence [6]. It is an efficient verifying mechanism but operates when the answer is generated. RAG-KG-IL [9] and Hyper-RAG [10] implement their verification functions earlier, but instead of the verification per se they enhance the process of obtaining the verified information in the knowledge storage [9].

Survey. Ni et al. [1] describes reliable RAG technology and categorizes this area according to reliability criteria but does not provide the authors' own framework.

TABLE I

SUMMARY OF THE REVIEWED LITERATURE (LATEST FIRST)

Ref.	Work	Year	Method	Main finding	Limitation
[2]	MARCH	2026	Multi-agent RL agents - Solvers, Proposers and Checkers	Reduced hallucination in RAG missions	High cost of training and computing
[6]	EAEV	2026	Counterfactual stability with entity-level witness alignment	Better hallucination detection in RAG assessments	Detects only

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Ref.	Work	Year	Method	Main finding	Limitation
[3]	FC-MAD	2026	Training-free multi-agent discussions with vote from the judge	High reliability and explainable fact-checking	High reliability and explainable fact-checking
[8]	A-ToM	2026	Adaptive Theory-of-Mind agent for coordination	Fewer coordination failures	Not designed for hallucination
[10]	Hyper-RAG	2026	Hypergraph-driven RAG enabling capture of multiple entity correlations	12.3% higher accuracy than direct use	LLM collected without on-line knowledge base
[4]	AgentNet	2025	Decentralized agents in a dynamic graph	Accuracy surpasses standard centralized solutions	reater cost of implementation depending on agent number
[5]	Darwish et al.	2025	LLM combined with rule-based logic and review loop	More consistent output	It relies on specialized rules
[9]	RAG-KG-IL	2025	Multi-agent RAG with knowledge graph and incremental learning	around 73% lesser hallucinations as compared to GPT-4o	Use of one health care case study for testing
[1]	Trustworthy RAG survey	2025	light survey according to reliability aspect	Provides a map to existing problems with regard to RAG	No new model introduced
[7]	Singhal et al.	2024	RAG with few-shot in-context learning	Fact checking based on supporting evidence and very little training	Limited by context window

IV. RESEARCH GAP

Reviewing literature gives several advantages and disadvantages presented below.

1. Many multi-agent techniques are too expensive, as they use reinforcement learning, several rounds of discussions, or numerous agents [2], [3], [4].
2. Most approaches do not work with the original answer in the sense that they are able to notice the mistake but correct it after it has already happened [2], [5], [6].
3. Some solutions are based on rules developed for a single domain [5].

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

4. There is no fast and universal verifier that can check retrieved evidence [1]. Knowledge graphs and hypergraphs can store and retrieve knowledge better [9], [10], but they do not check whether the information is appropriate, recent, and obtained from a reliable source.

The Main difference is presented in Fig. 3. The verifiers discussed in this paper are applied after the answer generation but the new verifier is utilized between retrieval and generation. Table II presents the comparison.

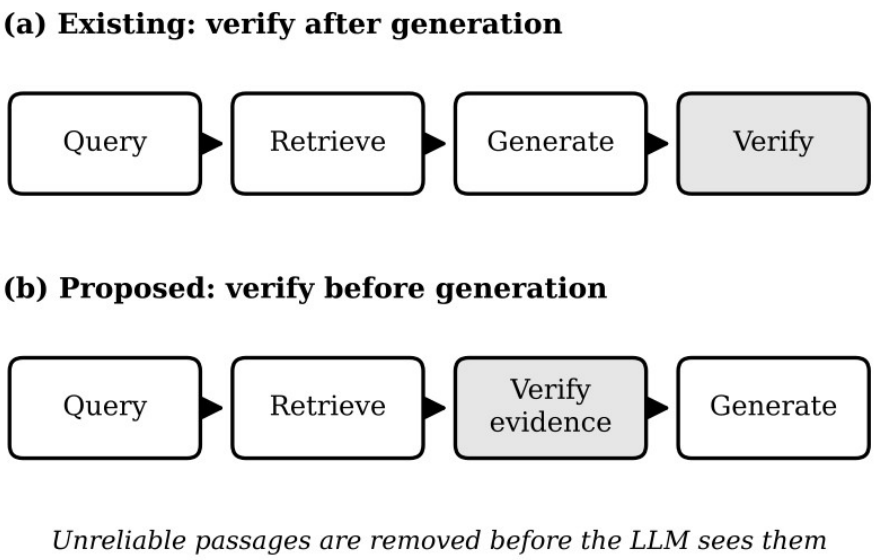


Fig. 3. Position of verification in (a) existing pipelines and (b) the proposed pipeline.

TABLE II

COMPARISON OF APPROACHES

Approach	Verification stage	Main cost
MARCH [2]	After generation, claim level	Reinforcement-learning training
EAEV [6]	After generation, entity level	Extra evidence-alignment step
FC-MAD [3]	Claim checking by debate	Several LLM calls and tokens
Darwish et al. [5]	After generation, review loop	Custom rules, review rounds
Singhal et al. [7]	Evidence extraction, then labelling	Extra LLM calls for evidence
RAG-KG-IL [9]	Knowledge-graph grounding of the answer	Graph building and updating
Hyper-RAG [10]	Hypergraph retrieval, no explicit check	Offline hypergraph construction

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Approach	Verification stage	Main cost
Proposed	Before generation	Small agents, one LLM call

V. PROPOSED METHODOLOGY

The concept involves introducing a minor verification layer between the retrieval agent and the generator. Instead of having a single major verifier, it employs three verification agents to handle different retrieved passages.

The agents are divided into three categories. The relevance agent determines how effectively a passage addresses the query. The consistency agent this evaluates the compatibility of various passages with each other and assigns lower scores to passages whose content is countered by many other passages. The freshness and source agent uses metadata like the publication date and type of source to apply penalties on older passages while rewarding trustworthy sources.

We start with a query q and a retrieved set called D = {d1, ..., dk}. The objective is to derive a subset called Ev from D so that the passages in Ev are relevant to the query q, are consistent among themselves, and are recent enough and from a trustworthy source. The LLM generates the answer a solely on the basis of query q and Ev.

How each agent can be built. The relevance agent applies a sentence embedding model to encode the query as well as each passage and use cosine similarity for scoring. The consistency agent runs a natural language inference model to assess whether the two passages contradict or not. The freshness agent observes the date of publishing and uses an exponential decay to give the passage a diminishing value with the time. The source agent associates the type of source with a corresponding reliability degree predetermined by the designer. Note that none of these agents needs any sophisticated LLM.

Scoring. Given a query q and a passage di (where i = 1, ..., k) each agent produces a score of 0 or 1:

$$r_i = \mathrm{sim}(q, d_i) \quad (1)$$

$$c_i = 1 - (1/(k-1)) \Sigma_{j \neq i} P_{\mathrm{contra}}(d_i, d_j) \quad (2)$$

$$f_i = \exp(-\lambda \Delta t_i) \quad (3)$$

where sim is the normalized similarity, Pcontra is the probability that two passages contradict each other, Δti is the age of the passage, and λ determines the rate of freshness decay. The reliability score s of the source is in the range from 0 to 1 based on the source type. The final score is

$$S_i = w_r r_i + w_c c_i + w_f f_i + w_s s_i, \text{ with } \Sigma w = 1 \quad (4)$$

and the verified evidence is

$$E_v = \text{Top-}m \{ d_i \mid S_i \geq \tau \} \quad (5)$$

where τ is the threshold and a maximum of m passages will be kept to stay within the context window. In Fig. 4, the four scores of a single passage are used to make the keep/drop decision.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

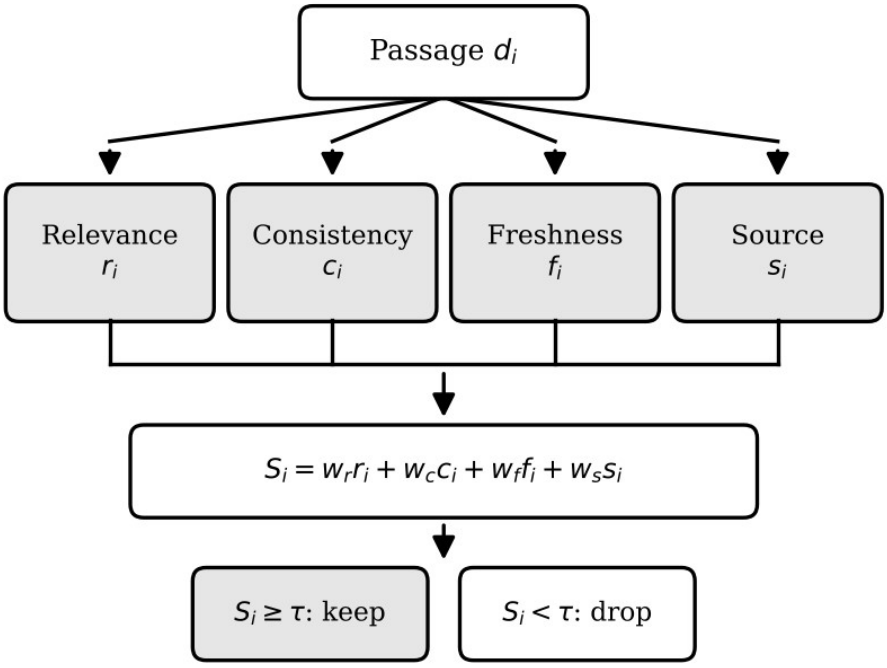


Fig. 4. Scoring and decision for a single retrieved passage.

Algorithm.

1. Obtain D ← Obtain(q, k).
2. For each d_i in D, execute the three agents in parallel to obtain r_i, c_i, f_i and s_i.
3. Calculate Si and create Ev.
4. If E_v is void, return "not enough evidence"
5. If not, invoke LLM once, a ← LLM(q, Ev), and return a along with Ev.

Walk-through. Suppose, for example, that a user inquires about the current refund rule for the company, and in the end three passages are returned: the current policy page, the old version of the policy from the past years, and the general FAQ entry. The relevance agent gives high scores to the policy page and FAQ entry while the score for the old version of the policy is lower because it tells about a rule that is no longer in force. At the same time, the consistency agent notices that the old version is contradicting the other two and this influences the score of the old policy negatively. The freshness agent reduces the score of the old version because of its age. Finally, the source agent gives the highest reliability score to the policy page. As a result of the weighting process, the old version is being thrown away as it does not accumulate enough points.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE III

PARAMETERS OF THE FRAMEWORK

Symbol	Meaning	Effect
k	Number of passages obtained	Use of larger k implies increased retrieval but also necessitates a longer checking time of the passages.
τ	Acceptance threshold	Greater limit is more exacting
m	Maximal number of passages retained	Indicates the length of the context
w	Agent weights	Establishes the relative importance of each verifier
λ	Freshness decay rate	More significant indicator leads to faster aging of the passages.

Example. When we choose k = 3 and weights of (0.4, 0.3, 0.2, 0.1), we take τ = 0.6. Let the scores be (0.9, 0.8, 0.9, 0.8) for d1, (0.7, 0.4, 0.5, 0.6) for d2 and (0.8, 0.9, 0.7, 0.9) for d3. Therefore S1 = 0.86, S2 = 0.56 and S3 = 0.82. As a result, d2 is disqualified and only d1 and d3 are seen by LLM. The provided values are supposed to be displayed on Fig. 5.

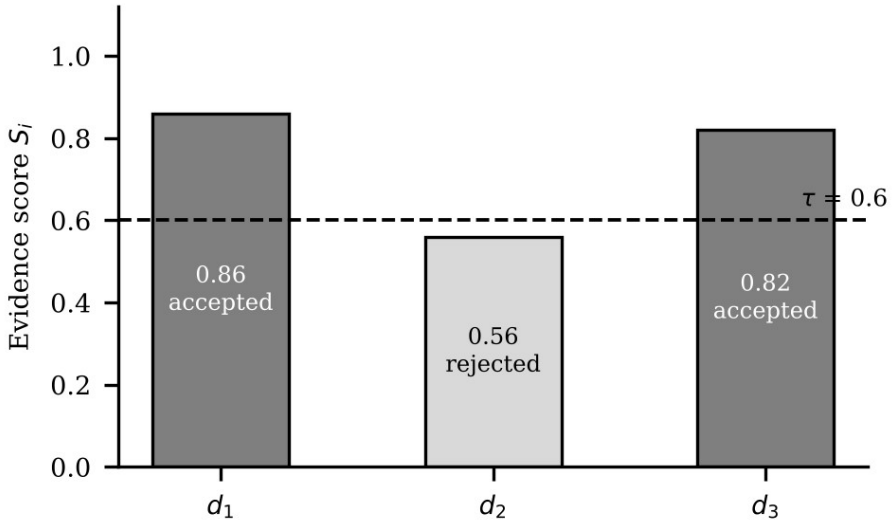


Fig. 5. Illustrative evidence scores of three passages against the threshold τ = 0.6.

Expense of the checks. Relevance checks and freshness checks entail just one check per passage. Consistency checking involves consideration of every possible pair and hence consists of roughly k(k−1)/2 comparisons, which remains small since k is around 5-10. The checks are independent from each other and can therefore proceed concurrently.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

VI. SYSTEM ARCHITECTURE

As shown in Fig. 6, the flow goes as follows. A query is sent to the retriever, and then three agents assess the retrieved passages, only the validated evidence being passed onto the LLM for presenting the final answer together with the evidence.

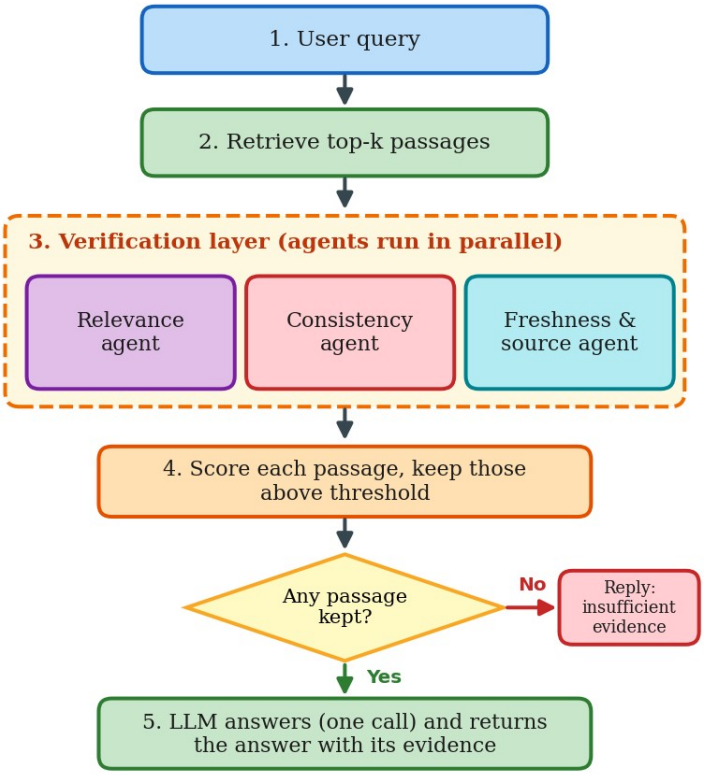


Fig. 6. Architecture and workflow of the proposed lightweight multi-agent verification framework.

Reason for being lightweight. Verification occurs prior to generating, thus eliminating any repetition of debate rounds. The agent comprises little elements, such as the similarity embedding, compact inference model, metadata rules, etc. and does not need to be a full LLM, allowing them to run in parallel. The large LLM will only be called once, in a summarized and shortened context. If Cag is the cost of agents and n Cverify is the cost of n verification rounds post-generation, the value we aim to achieve is Cag ≪ n Cverify.

Possible tools. The framework may be established using several open-source components, which are listed in Table IV.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE IV

CANDIDATE TOOLS FOR IMPLEMENTATION

Component	Possible choice
Language	Python
RAG framework	LangChain or LlamaIndex
Retrieval	Sentence embeddings with FAISS or Chroma
Consistency check	Compact NLI model (Hugging Face)
Generator	Open-source or API-based LLM
Test data	FEVER, HotpotQA, Natural Questions, TruthfulQA

VII. RELEVANCE TO CURRENT TRENDS

The concept of hallucination control is a hot topic, being discussed in contexts very far apart, ranging from claim-level multi-agent verification [2], entity-level verification [6], explainable dialogue [3] agent coordination [4], [8], knowledge graph/hypergraph search [9], [10].Businesses are using large language model assistants as well to process their internal documentation and support data, where errors lead to loss of money or trust. This explains the growing interest in such trending topics as Trustworthy AI, explanation generation, verification layers, to meet the need for a cheap solution.

VIII. APPLICATIONS AND ADVANTAGES

The verification layer for RAG helps in answering queries accurately in different fields. In the sphere of medicine, a healthcare virtual assistant answering per clinical rules should utilize the latest guideline. In the sphere of education, tutoring services based on the course content have to provide relevant answers. In the spheres of finance and law, due to frequent changes to regulations and contradictions in documents it is important to perform freshness and consistency checks. Enterprise chatbots and customer support operate with various quality internal documents. Facts verification systems can also use the only relevant evidence [3], [7].

TABLE V

EXAMPLE APPLICATION AREAS

Area	Main risk	Most useful check
Medical profession	Obsolete recommendations	Currency and origin of information
Education	Incorrect educational content	Applicability
Finance and law	Altered or contradictory rules	Uniformity and timeliness
Customer support	Contradictory document quality	Origin and applicability

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

There are several key benefits. First, unreliable content is filtered out ahead of time as opposed to downstream. The additional computational cost remains low, because agents are light and LLM is called only once. The architecture is modular so that an agent can be replaced in a specific domain. The answer is always provided with evidence.

IX. ETHICAL AND SOCIETAL CONSIDERATIONS

Presenting proof for an answer contributes to increasing system transparency, while remaining silent when the proof is inadequate seems to be a safer option than giving a false but convincing answer. The source reliability score may introduce bias when genuine sources always receive low scores, meaning the scoring system should be fine-tuned accordingly. Knowledge sources may include sensitive information, which means access policy is needed to treat the resulting text as unreliable evidence. A trusted answer or a verified one can still be incorrect, so users should be deprived of false confidence that the answer is correct. Moreover, the light design consumes less processing power and energy in each case compared to multi-round verification.

X. PROPOSED EVALUATION

Comparison should occur between the framework, plain RAG, and, at least, one post-generation verifier. We propose a comparison based on the following measures:

- Factual accuracy and hallucination rate, with hallucination rate meaning the proportion of answers that contains some unsupported claim;
- Faithfulness of answer with respect to verified evidence;
- Latency and token consumption for each query;
- Abstention rate, meaning the proportion of evidence that was impacted by the filter and hence would not be considered as too informative.

TABLE VI

PLANNED EVALUATION MEASURES

Measure	What it shows
Rate of hallucination	Proportion of responses containing unsubstantiated claims
Trustworthiness	Quality of responses based on established evidence
Correctness of response	Accuracy rate in relation to reference answers
Speed of response	Time dedicated to each inquiry
Number of tokens used	Volume of token transferred in and out of the LLM
Rate of non-response	Frequency of rejections by the system

Comparison will need to cover three cases: pure RAG, RAG with a post-generation verifier, and the proposed framework. Ablation experiments with only one agent removed will quantify the contribution of individual checks, while varying k and τ will help understand sensitivity of the results. The experiments will need to keep retriever, embedding model, and LLM identical across different frameworks, so that only verification stage will differ. Each of the experiments will need to be repeated with different seeds or

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

settings, and results will be reported with their dispersion. The data to utilize will include FEVER-like factual-checking sets, HotpotQA, Natural Questions, TruthfulQA, and one more dataset of the specific domain.

XI. DISCUSSION, LIMITATIONS AND FUTURE WORK

Discussion. The analyzed study shows a balance between verification strength and cost. The method that utilizes a few agents or discussions may give accurate results with good transparency, but it has to be performed in a cumbersome manner. The proposed method sacrifices precision for efficiency but allows conducting verification at early stages. It has nothing to do with post-generation methods such as MARCH or EAEV by means of doing it simultaneously with these methods so that early verification is capable of filtering out the gross mistakes. Also, the method can be used before the knowledge-graph and hypergraph retrievers like RAG-KG-IL [9] and Hyper-RAG [10] so that they can run only valuable queries at the outset.

Limitations. The framework developed is theoretical, thus hasn't been tried in practice. It relies on retrieval, so the verification cannot help if the needed passaging hasn't been retrieved. The values of τ, λ, and the weights would require tuning and do not necessarily transfer from domain to domain. Too strict filtering may cut useful evidence, while absent dates or source tags diminish freshness and source agent's efficiency. Lastly, validating the evidence prior to generation does not prevent the LLM from making an internal hallucination.

Future work. The plans for the next steps include building the prototype, learning the weights and thresholds instead of setting them by hand, activating additional agents for only high-risk requests, testing on medical, law, and education data, and implementing light checks after generation for all the remaining cases.

XII. CONCLUSION

LLMs cannot be trusted in most significant use cases because of hallucination, and RAG does not eliminate this problem since retrieval is not foolproof in several respects. The papers we examined indicate that validation is effective, but advanced techniques are expensive and mainly have an effect post-generation. On the other hand, we proposed a cost-efficient multi-agent framework that evaluates retrieved documents in terms of relevancy, coherence, freshness, and source trustworthiness prior to the generation process, retaining only trustworthy pieces of evidence and making a single petition to LLMs post-evaluation. This solution is aimed at increasing trustworthiness while having little impact on the budget level but still needs to be validated in practice with the help of appropriate testing and assessment. Provided that the initial filter works as planned, it can be integrated into the already existing RAG solutions with little effort and later combined with post-generation operational checks for sensitive applications.

REFERENCES

[1] B. Ni *et al.*, "Towards trustworthy retrieval augmented generation for large language models: A survey," arXiv:2502.06872, Feb. 2025.

[2] Z. Li, Y. Zhang, P. Cheng, J. Song, M. Zhou, H. Li, S. Hu, Y. Qin, E. Zhao, X. Jiang, and G. Jiang, "MARCH: Multi-agent reinforced check for hallucination," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Jul. 2026, pp. 39389–39415.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

[3] T.-A. Nguyen, B. Vo, and T. K. Tran, "Debating to verify: A robust and explainable multi-agent LLM system for fact-checking," *ICT Express*, May 2026, doi: 10.1016/j.icte.2026.05.017.

[4] Y. Yang, H. Chai, S. Shao, Y. Song, S. Qi, R. Rui, and W. Zhang, "AgentNet: Decentralized evolutionary coordination for LLM-based multi-agent systems," in *Proc. 39th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

[5] A. M. Darwish, E. A. Rashed, and G. Khoriba, "Mitigating LLM hallucinations using a multi-agent framework," *Information*, vol. 16, no. 7, p. 517, Jun. 2025, doi: 10.3390/info16070517.

[6] R. Jia, Z. Fang, M. Wu, J. Lu, and Y. Zhang, "Evidence-aligned entity verification for hallucination detection in retrieval-augmented generation," in *Findings Assoc. Comput. Linguistics: ACL 2026*, Jul. 2026, pp. 29538–29554.

[7] R. Singhal, P. Patwa, P. Patwa, A. Chadha, and A. Das, "Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs," in *Proc. 7th Fact Extraction and VERification Workshop (FEVER)*, Nov. 2024, pp. 91–98.

[8] C. Mu, Y. Zeng, Q. Zhang, K. Shao, C. Chu, H. Guo, D. Jia, Z. Wang, and S. Hu, "Adaptive Theory of Mind for LLM-based multi-agent coordination," in *Proc. 40th AAAI Conf. Artif. Intell. (AAAI)*, 2026, pp. 29608–29616.

[9] H. Q. Yu and F. McQuade, "RAG-KG-IL: A multi-agent hybrid framework for reducing hallucinations and enhancing LLM reasoning through RAG and incremental knowledge graph learning," arXiv:2503.13514, Mar. 2025.

[10] Y. Feng et al., "Hyper-RAG: Combating LLM hallucinations using hypergraph-driven retrieval-augmented generation," *Nature Commun.*, vol. 17, Art. no. 5778, Apr. 2026, doi: 10.1038/s41467-026-71411-1.