 ISSN (Online): 3140-5312

A Self-RAG-Based Triple Factuality Verification Pipeline for Local LLM KARMA

Seung-gyeom KIM[1], Giseop Noh[2]

[1]BE Course, Department of Software and Communications Engineering, Hongik University, South Korea

[2]Professor, Department of Software and Communications Engineering, Hongik University, South Korea

Corresponding author: Giseop Noh (kafa46@hongik.ac.kr)

Abstract

Large language models (LLMs) can extract entities and relations from unstructured documents to construct or expand knowledge graphs. However, generated triples may be insufficiently grounded in the source text or contain incorrect relations derived from a model's internal knowledge. KARMA is a knowledge graph enrichment framework in which multiple LLM agents perform document analysis, entity and relation extraction, schema alignment, and knowledge integration. To reduce dependence on commercial services, this study implements KARMA with a local LLM and combines it with Self-RAG retrieval and self-evaluation to assess external evidence. The pipeline consists of document preprocessing and semantic chunking, chunk-level evidence evaluation, candidate triple extraction and source-based quality checking, triple-level factuality verification, and knowledge graph acceptance. In a preliminary experiment, an NLP paper was divided into 14 chunks, from which 85 initial triples were extracted; 28 passed the quality check. External documents confirmed the head entity for six triples, but no complete claim including the relation and tail entity was verified. Consequently, no triple was automatically accepted. The results reveal limitations in the temporal coverage of the 2018 Wikipedia corpus and in filtering non-body PDF content, motivating more recent retrieval corpora, improved preprocessing, and review procedures for potentially novel facts.

Keywords: Knowledge Graph, Local LLM, KARMA, Self-RAG, Factuality Verification

1. Introduction

Advances in LLM have increased interest in knowledge graphs as a means of extracting structured knowledge from unstructured data and supporting reasoning over relationships between entities [1]. Conventional knowledge graph construction often relies on domain experts who manually analyze documents or on rule-based natural language processing systems. Although these approaches can provide relatively high reliability, they require substantial time and labor as the volume of documents increases [2].

KARMA addresses this limitation through nine commercial-LLM-based agents that collaboratively and iteratively analyze unstructured documents, extract and validate triples in the form of head entity, relation, and tail entity, and integrate them into a knowledge graph [3]. The agents divide responsibilities across document normalization, entity and relation extraction, schema alignment, knowledge validation, and conflict resolution.

To reduce API costs in large-scale literature processing, this study replaces the commercial LLMs in KARMA with a local LLM. Response preprocessing and a domain-adaptive schema induction mechanism are also used to mitigate unstable triple output, duplicate entities, and fragmented relation expressions. However, these improvements primarily target the formal and structural quality of extracted results and do not independently verify whether generated triples are factually supported by the source text and external knowledge.

ISSN (Online): 3140-5312

Therefore, this study applies the Self-RAG framework to strengthen evidence-based verification of triples extracted by local LLM KARMA. Input papers or documents are divided into semantically coherent chunks based on paragraph, sentence, and section boundaries to minimize context loss. The complete pipeline is designed to determine whether external retrieval is necessary before retrieving relevant documents and assess their relevance to the input. Candidate triples extracted from the chunks are checked against both the source text and external documents. Finally, the preceding evaluations are combined with knowledge graph acceptance criteria to determine whether each triple should be integrated.

The objectives of this study are threefold. First, it analyzes the structural characteristics of KARMA and Self-RAG. Second, it proposes a local LLM KARMA pipeline that incorporates Self-RAG-based retrieval and self-evaluation, while designing adaptive retrieval as a component for later end-to-end integration. Third, it analyzes limitations of the retriever, preprocessing procedure, and evaluation model identified in the preliminary experiment and presents directions for further improvement.

2. Background

2.1 KARMA

KARMA is a multi-agent framework that automatically enriches an existing knowledge graph by extracting new knowledge from unstructured documents [3]. Knowledge graphs generally represent facts as triples consisting of a head entity, a relation, and a tail entity [1]. For example, the statement "Self-RAG uses adaptive retrieval" can be represented as follows

(Self-RAG, uses, adaptive retrieval)

A key characteristic of KARMA is that no single LLM performs the entire process from document analysis to final knowledge integration. Multiple agents divide the work, and each downstream agent reviews the output produced in the preceding stage. KARMA includes entity discovery, relation extraction, schema alignment, and conflict resolution. These stages help ensure that newly extracted entities and relations conform to a domain-specific knowledge graph schema and reduce the integration of information that logically conflicts with existing knowledge or has low confidence [3].

The original KARMA employs high-performing commercial GPT models so that individual agents can follow complex instructions and review their outputs [3]. In a local environment, however, model size, GPU memory, and inference speed are constrained. Furthermore, when an error propagates to downstream agents, an incorrect triple may remain in the final output despite multiple review stages.

Accordingly, local LLM KARMA requires not only the preservation of the original agent structure but also a verification process that directly compares generated triples with the source text or external documents. This study applies Self-RAG retrieval-necessity assessment and evidence-based self-evaluation to that verification process.

2.2 Conventional RAG and Self-RAG

Retrieval-Augmented Generation (RAG) enables a language model to retrieve external documents related to an input and use them during generation rather than relying only on knowledge acquired during training [4]. This approach can reduce factual errors by grounding a response in external information when the model's internal knowledge is insufficient or inaccurate. However, conventional RAG commonly retrieves a fixed number of documents without adequately determining whether retrieval is necessary or whether the retrieved documents are relevant to the input [5].

Table 1. Four types of reflection tokens in Self-RAG

Type	Definitions	Output
Retrieve	Is external retrieval necessary?	{yes, no, continue}
ISREL	Is the retrieved document relevant to the input?	{relevant, irrelevant}
ISSUP	Is the generated output supported by the retrieved document?	{fully supported, partially supported, no support}
ISUSE	Is the final output useful to the user?	{5, 4, 3, 2, 1}

Self-RAG addresses these limitations by allowing a language model to retrieve external documents only when needed, generate an output using the retrieved evidence, and evaluate the quality of that output [5]. In addition to

ISSN (Online): 3140-5312

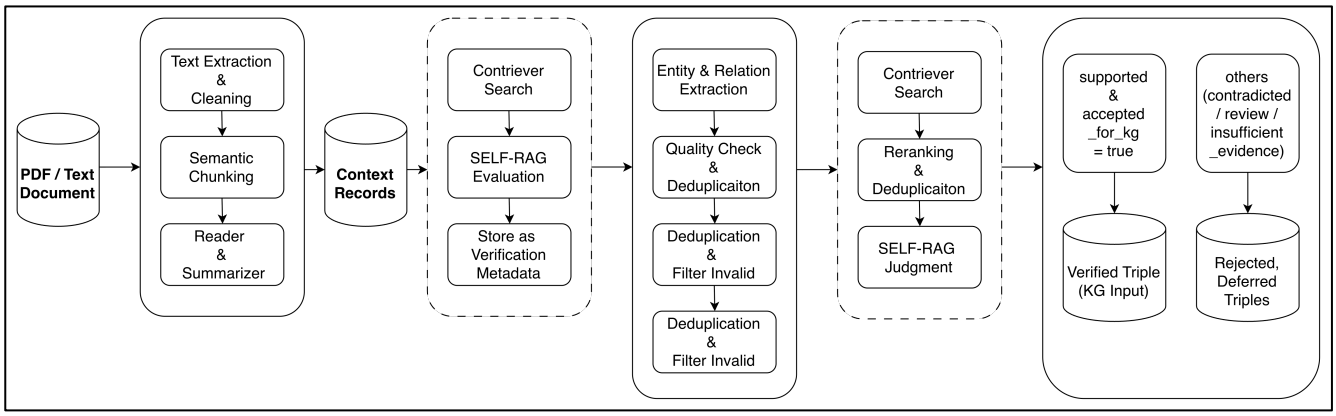

Fig. 1. Flowchart of the Self-RAG-Based Local LLM KARMA Pipeline

ordinary output tokens, the model generates reflection tokens that represent its decisions. These tokens assess whether external retrieval is necessary, whether a retrieved document is relevant to the input, how strongly the generated output is supported by the document, and whether the final output is useful. Self-RAG therefore adaptively determines retrieval necessity instead of retrieving the same number of documents for every input, and it progressively evaluates retrieved evidence and generated content before selecting the final output. The reflection token types are summarized in Table 1 [5].

3. Methods

3.1 Self-RAG-Based Local LLM KARMA

The proposed pipeline comprises document preprocessing and semantic segmentation, chunk-level external evidence evaluation, KARMA-based candidate triple extraction, triple-level factuality verification, and generation of verified knowledge graph input. Both the language model and the retriever run locally without relying on commercial external APIs.

First, the main text is extracted from a PDF or text document and divided into semantic chunks according to paragraph, sentence, and section boundaries. Sections with low relevance to the main content, such as references and acknowledgments, are excluded through section classification. KARMA's Reader then selects passages relevant to knowledge extraction from each chunk, and the Summarizer condenses them to produce a TARGET for triple extraction. The source passage corresponding to each TARGET is retained separately to verify the extraction result. The preceding and following chunks are also attached as auxiliary context to reduce context discontinuity. However, entities and relations are extracted only from the TARGET to prevent surrounding context from being incorrectly treated as new knowledge. External evidence retrieval uses a corpus built from a 2018 Wikipedia snapshot split into passages of approximately 100 words. In this study, a retrieved document refers to one such passage rather than an entire Wikipedia article. Contriever encodes each TARGET and passage as numerical vectors and computes semantic similarity between them [6]. Passage vectors are stored in a FAISS index. FAISS is a vector similarity search library that efficiently finds items near a query vector in a large vector collection; the index stores search vectors rather than the Wikipedia source text itself [7]. Each preprocessed chunk undergoes Self-RAG-based external evidence evaluation using Contriever and the Wikipedia FAISS index. Contriever returns the top five passages for each TARGET, and Self-RAG evaluates their relevance to and support for the TARGET. A passage is accepted as external evidence only when both the relevance and support scores are at least 0.5 and Self-RAG classifies it as fully supporting the TARGET. The system also checks whether at least 80% of the TARGET's key content words are preserved and whether major entity names and numerical values remain intact. The current implementation retrieves evidence for every chunk without separately predicting retrieval necessity. A chunk is not removed even when no external evidence is accepted. Evaluation results and retrieved passages are stored as verification metadata and passed to the triple extraction stage.

During triple extraction, KARMA's Entity Extraction Agent, Relationship Extraction Agent, and Schema Alignment Agent are executed. These agents extract entities and relations from the TARGET and generate candidate triples after normalizing entity aliases and relation expressions. Each generated triple is checked against the retained source passage for the presence of its head and tail entities, the correspondence of its relation expression, and structural completeness. Results are removed when their components cannot be found in the

 ISSN (Online): 3140-5312

Table 2. Results of the pipeline

Metric	Measurement	Result
External Evidence Acceptance Rate	Accepted evidence chunks / Total chunks	0.000
Triple Quality Pass Rate	Valid triples / Initially extracted triples	0.329
Head Evidence Coverage	Triples with head evidence / Valid triples	0.214
Complete Claim Evidence Rate	Triples with complete claim evidence / Valid triples	0.000
Automatic KG Acceptance Rate	Automatically accepted triples / Valid triples	0.000

source, when the relation is excessively long, or when the relation is generated as a full sentence. Ambiguous results are assigned for further review. The original KARMA Conflict Resolution Agent and Evaluator Agent are not directly used in the current integrated pipeline; instead, the subsequent rule-based quality check and Self-RAG verification determine final acceptance.

Triples that pass the quality check are converted into natural-language verification queries. For each query, Contriever retrieves the top ten passages from the same Wikipedia FAISS index. The results are reranked using not only initial embedding similarity but also the presence of head aliases, relation expressions, and the tail entity. Passages that do not correspond to the head entity and near-duplicate passages are removed. Up to three passages are then selected as final evidence from a maximum of five candidates.

During triple verification, a rule-based check first determines whether the head entity, relation, and tail entity can be identified in the retrieved passages. If no passage corresponding to the head entity is found, the triple is labeled insufficient_evidence. When verifiable evidence is available, Self-RAG relevance and support assessments are additionally applied. The integrated decision labels a triple as supported when external evidence sufficiently supports it, contradicted when an explicitly conflicting fact is identified, review when partial evidence or conflicting judgments require further examination, and insufficient_evidence when sufficient evidence cannot be found.

Only triples labeled supported and satisfying accepted_for_kg=true are stored as verified knowledge graph input. Triples labeled contradicted are rejected, whereas review and insufficient_evidence cases are retained with their decisions and retrieved evidence for subsequent examination. By separating chunk-level evidence evaluation before extraction from triple-level factuality verification after extraction, the proposed pipeline preserves source information while automatically accepting only triples with sufficient evidence.

4. Experimental Results and Analysis

This study conducted an experiment on a natural language processing paper to examine the initial behavior and performance of the integrated Self-RAG-based local LLM KARMA pipeline. The experiment was conducted before integration of Reader and Summarizer preprocessing, and the semantically segmented source chunks were therefore used directly as TARGETs. The purpose was to first verify the operation of the core sequence comprising chunk-level evidence evaluation, KARMA-based triple extraction, triple quality checking, and external factuality verification. In addition, the experiment used unconditional retrieval, meaning that external evidence was retrieved for all 14 chunks without applying the retrieval-necessity classifier. Rule-based chunking based on paragraph, sentence, and section boundaries divided the input into 14 chunks, from which KARMA extracted 85 initial triples. The overall results are shown in Table 2.

The chunk-level external evidence acceptance rate is the proportion of chunks for which retrieved passages sufficiently support the TARGET and are accepted as verification evidence. The external corpus consisted of Wikipedia passages of approximately 100 words. Self-RAG evaluated the top five passages returned by Contriever for each TARGET. A passage was accepted only when both relevance and support scores were at least 0.5, Self-RAG labeled it as fully supporting the TARGET, at least 80% of the TARGET's key content words were preserved, and major entity names and numerical values were retained. A total of 70 evidence candidates were retrieved for the 14 chunks, but none was accepted. Self-RAG classified seven chunks as insufficiently relevant to the retrieved passages. The remaining seven achieved

ISSN (Online): 3140-5312

relevance scores of at least 0.5 but failed the support threshold and were judged not to sufficiently support the TARGET or to potentially conflict with it. Consequently, the chunk-level external evidence acceptance rate was 0.000. This result does not indicate that retrieval returned no passages; rather, none of the retrieved passages met the evidence acceptance criteria.

The triple quality pass rate is the proportion of initial triples whose head, relation, and tail are structurally complete and correspond to the source text. Of 85 initial triples, 28 passed the quality check, yielding 0.329. Among the remaining triples, 34 were classified as invalid and 23 as requiring further review. The main causes were a head entity not found in the source text, an insufficient correspondence between the relation and the source expression, and ambiguity requiring additional interpretation. Because some triples contained more than one issue, the sum of issue counts does not equal the number of excluded triples.

The head entity evidence coverage is the proportion of quality-approved triples for which the head entity was found in an external passage. Both the original entity expression and its abbreviations and aliases were considered. The head entity was confirmed for 6 of the 28 valid triples, producing a coverage of 0.214. No directly corresponding external evidence was found for the remaining 22 triples.

The complete claim evidence rate is the proportion of valid triples for which the head, relation, and tail were all confirmed in external passages. Of the six triples with confirmed head entities, four lacked an expression supporting the relation, and the remaining two did not provide evidence for the entire claim including the tail entity. Thus, none of the 28 valid triples had complete claim evidence, resulting in a rate of 0.000. This finding shows that identifying a head entity in an external document alone is insufficient to establish the factuality of the complete triple.

The knowledge graph automatic acceptance rate is the proportion of valid triples that were sufficiently supported by external evidence and accepted as final knowledge graph input. Of the 28 valid triples, 22 were labeled insufficient_evidence by the rule-based step because their head entities were not identified in the retrieved passages. The remaining six also lacked sufficient Self-RAG evidence for the complete fact, including the relation and tail entity. Consequently, all 28 triples were classified as insufficient_evidence, and none was automatically accepted, yielding an automatic acceptance rate of 0.000. Importantly, insufficient_evidence does not mean that a triple is false. It indicates that the current external corpus did not provide sufficient evidence to determine its factuality. Therefore, unverified triples were not immediately discarded; their decisions and retrieved evidence were retained for later review.

The preliminary experiment also revealed that non-body content, including author names and paper titles in the references as well as table titles and captions, was recognized as entities and used as the head or tail of generated triples. This likely occurred because source chunks were passed directly to triple extraction without prior selection of relevant passages. To mitigate this issue, the subsequent pipeline added a preprocessing stage in which the KARMA Reader selects passages relevant to knowledge extraction and the Summarizer organizes them into TARGETs. Because the complete experiment has not yet been repeated with the Reader and Summarizer enabled, further experiments are needed to quantify improvements in triple quality and external evidence acceptance.

5. Conclusion and Future Work

5.1 Conclusion

This study proposed an integrated pipeline combining Self-RAG retrieval and self-evaluation to verify the factuality of triples generated by local LLM KARMA. The pipeline comprises semantic segmentation of input documents, chunk-level external evidence evaluation, KARMA-based candidate triple extraction, source-based quality checking, triple-level factuality verification, and generation of verified knowledge graph input. The complete design includes adaptive retrieval based on retrieval-necessity assessment. However, because this module had not yet been integrated into the end-to-end pipeline, the preliminary experiment used unconditional retrieval and searched external evidence for every evaluation target.

In the preliminary experiment, the source document was divided into 14 semantic chunks and used directly as TARGETs without Reader and Summarizer preprocessing. KARMA extracted 85 initial triples, of which 28 passed the structural completeness and source-grounding quality criteria. The chunk-level evaluation returned 70 external passages, but none met the evidence acceptance criteria. Only six of the 28 valid triples had a head entity identified in external evidence, and none had evidence for the complete fact including its relation and tail. Accordingly, all 28 triples were labeled insufficient_evidence, and none was automatically accepted into the knowledge graph. These results

ISSN (Online): 3140-5312

demonstrate that the current conservative acceptance policy prevents automatic integration of triples lacking sufficient external evidence.

The experiment also exposed limitations in the temporal coverage of the retrieval corpus and in document preprocessing. The corpus was built from a 2018 Wikipedia snapshot, whereas the input paper was published in 2025; therefore, entities and findings introduced after 2018 may not be represented in the corpus. In addition, non-body information such as reference titles and author names, table titles, and captions was recognized as entities and used in generated triples. Because no controlled comparison was conducted to measure the effects of corpus age and preprocessing quality on evidence acceptance, these factors should be interpreted as constraints of the current experimental setting rather than direct causes.

The proposed pipeline operated as a verification mechanism that did not automatically integrate triples lacking sufficient evidence. However, the current experiment cannot determine whether the low acceptance rate primarily resulted from the temporal limitations of the corpus, retrieval performance, document preprocessing errors, or stringent acceptance thresholds. Future work should therefore isolate and evaluate the effect of each factor.

5.2 Future Work

Future experiments should expand the temporal coverage of the retrieval corpus by using up-to-date Wikipedia documents. To objectively evaluate the effect of corpus construction time, comparative experiments should use the 2018 Wikipedia snapshot, a recent Wikipedia snapshot, and recent scholarly literature, and analyze changes in the chunk-level external evidence acceptance rate and complete claim evidence rate.

Preprocessing must also be improved to reliably select valid body text from unstructured documents. The current implementation identifies some non-body sections using section headings and string patterns; however, if PDF conversion destroys layout and paragraph boundaries, references, table titles, and captions are difficult to distinguish from body text. A layout-aware preprocessing method should therefore use positional, font, and page-layout information to separate titles, author information, headers and footers, references, tables, and captions. The full pipeline should also be rerun with the Reader's passage selection and the Summarizer's summarization functions added after the preliminary experiment, and the reduction in non-body entities and invalid triples should be evaluated.

Because the integrated experiment retrieved evidence for every evaluation target, the separately tested retrieval-necessity classifier should be connected to the chunk and triple verification stages. Adaptive retrieval should then be compared with unconditional retrieval in terms of reductions in unnecessary retrieval, omission of required evidence, external evidence acceptance, and final verification performance.

Finally, triples labeled insufficient_evidence may represent not only false information but also novel facts absent from the existing corpus. Such triples should not be immediately discarded. Instead, a separate review process should determine their eligibility for knowledge graph integration by consulting recent scholarly literature, domain-specific databases, or domain experts.

Conflict of Interest

The authors declare that there are no potential conflicts of interest related to this paper.

Acknowledgement

This research was supported by the Ministry of Science and ICT(MSIT), Korea, under the Sejong SW Convergence Cluster 2.0 Project supervised by the National IT Industry Promotion Agency(NIPA) (Grant No. PJTS080126041007000100030020 0).

References

[1] B. Zhang and H. Soh, “Extract, Define, Canonicalize: An LLM-based framework for knowledge graph construction,” in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, pp. 9820–9836, 2024, doi: 10.18653/v1/2024.emnlp-main.548.

[2] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, “A survey on knowledge graphs: Representation, acquisition, and applications,” IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494–514, 2022, doi: 10.1109/TNNLS.2021.3070843.

[3] Y. Lu, W. Wu, X. Zhao, R. Peng, and J. Wang, “KARMA: Leveraging multi-agent LLMs for

 ISSN (Online): 3140-5312

automated knowledge graph enrichment," in Advances in Neural Information Processing Systems 38, 2025.
[4] P. Lewis et al., "Retrieval-Augmented Generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
[5] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in The Twelfth International Conference on Learning Representations, 2024.
[6] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised dense information retrieval with contrastive learning," Transactions on Machine Learning Research, 2022.
[7] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.

Author Information

Seung-gyeom Kim
2022~ now: BE, Dept. of Software & Communications Engineering, Hongik University
Email: ksg072896@naver.com
Research Interest: LLM, NLP, RAG

Giseop Noh
1998: BE, Engineering, Korea Air Force Academy
2009: MS, Computer Science, University of Colorado, USA
2014: PhD, Computer Engineering, Seoul National University
2016.12~2018.02: Assistant Professor, Dept. of Computer & Information Science, Korea Air Force Academy
2018.03~2025.02: Assistant Professor, Dept. of Artificial Intelligence Software, Cheongju University
2025.03~now: Assistant Professor, Dept. of Software & Communications Engineering, Hongik University
Email: kafa46@hongik.ac.kr
Research Interest: NLP, Neural Network Theory, Reinforcement Learning