An Adaptive RAG-Based Question-Answering System in the Context of Industry 5.0

Dr. Krishnendu Mukherjee[1*]

[1]Associate Professor, School of Artificial Intelligence, Sai University, Tamil Nadu.

*Corresponding Email: kmukherjeemech@gmail.com

Abstract

In this paper, a pragmatic literature review approach has been shown to select research papers to determine important technologies in the context of Industry 5.0 or I 5.0 such as Artificial Intelligence (AI), Internet of Things (IOT), Collaborative Robot (Cobot), Cyber Physical System (CPS), Human Machine Interface (HMI), Edge Computing, Big Data, Digital Twin, Virtual Reality (VR), Reinforcement Learning (RL), Large Language Model (LLM), Multiple Criteria Decision Analysis (MCDA) etc. The crux of this paper is to develop an economical Adaptive RAG-based (ARAG) system which could generate contextual relevant responses to a user's query. A two-stage hybrid zero-resource hallucination detection system has been developed to detect hallucinations in the generated response. A binary classifier has been developed using Mistral 7B for fact checking against reliable resources and a panel of multiple Large Language Models (LLMs), namely, Mistral 7B, Llama 3 8B, and Llama 2 7B, has been used to evaluate the factual accuracy of the generated response asynchronously using the 5-point Agreement Scale. Mistral 7B has shown a very high correlation with human judges. Open source or free resources are used to develop the ARAG, and, thus, the ARAG is economical. A brief discussion on multilingual responses is also included.

Keywords: Large language model; Adaptive rag; Industry 5.0; Rag; Mistral; Hallucination.

1. INTRODUCTION

"*Degrowth is not the sign of fail but stability. Degrowth does not mean lower profitability. Degrowth does not waste resources. Degrowth does not destabilize the system. Degrowth does not represent scarcity but enough" – Michael Rada*

Industry 5.0 is not a mere extension of Industry 4.0 or the human-centric application of manufacturing and production processes. It is the skeptical use of man, machine, material, resources, and technology (3MRT). It is giving utmost priority to the ethical use of artificial intelligence, data privacy, collaborative environment of man and machine, integrated use of lean and agile systems to develop a resilient supply chain to combat uncertainty, and leveraging big data, cloud, and other technologies without sacrificing the value of life and the environment. In Industry 4.0, advanced technologies have been used extensively. In Industry 5.0, on the other hand, social, environmental, and economical aspects are considered heavily to have a better future considering the compelling effects of global climate, COVID, Ukraine-Russia war, the humanitarian catastrophe in Gaza etc. Industry 5.0 demands the judicious utilization of man, machine, material, resources, and technology. According to the European Commission's agenda, Industry 5.0 prioritizes eco-friendly, human-centric, and resilient supply chains (Breque *et* al. [6]). Industry 5.0 represents a socially pulled and a technologically pushed digital transformation phenomenon (Ghobakhloo *et* al. [11]).

Ghobakhloo *et* al. [11] further showed the architectural design of Industry 5.0 which consists of the following components:

1. Enabling Technologies of Industry 5.0:

   A. Facilitating Technologies: Additive Manufacturing, Networking and Communication Technologies, Embedded System, Enterprise Systems, Internet of Everything, Industrial Control System, Machine Learning and Cognitive Computing, Blockchain, Cloud and Edge Computing, and Big Data Analytics.

   B. Emerging Technologies: Cognitive Cyber-Physical Systems, Cognitive Artificial Intelligence, Human Interactions and Recognition Technologies, Extended Reality, Industrial Smart Wearable, Intelligent Robots, Intelligent Energy Management System, Multiscale Dynamic Modelling and Simulation, and Smart Product Life Cycle Management.

2. Techno-functional Principles of Industry 5.0: Decentralization, Vertical Integration, Horizontal Integration, Virtualization etc.

3. Industry 5.0 Components: Smart Customers, Smart Logistics, Smart Products, Smart Stakeholder, and Smart Supply Chain.

4. Economic Bottom Line, Environmental Bottom Line, and Social Bottom Line.

Human-Machine Collaboration (HMI), one of the most important features of Industry 5.0, is basically a bidirectional learning process and a reciprocal collaboration between human and machine to perform shared tasks effectively (Ansari *et* al. [4]). Artificial Technology (AI) is the enabling technology of Industry 5.0 and active learning, explainable AI, simulated reality, conversational interfaces, and security are the enabling technologies of human-centric AI (Rožanec *et* al. [17]). A conversational interface leverages AI and LLM to give better interactions between human and machine. It could be basic-chatbots, text-based assistants, and voice-based assistants. LLMs are sophisticated computational models with linguistic intelligence to comprehend and generate human language to answer the user's query (Vlacic *et* al. [23], Kasneci *et* al. [14], and Chang *et* al. [7]). Human-centric technological innovations are the fundamental blocks of I5.0 (Ghobakhloo *et* al. [11]). The user sends queries to the machine through the conversational interface and the machine answers to the user's query so that man and machine can perform a shared task effectively. LLM plays a dominating role in such applications. LLMs such as ChatGPT, DALLE etc. are playing the pivotal role in transforming manufacturing and industrial processes by enhancing operational efficiency, reducing downtime, and accruing economic savings through predictive maintenance and real-time data analysis (Sai *et* al. [19]). Kiangala and Wang [15] developed a chatbot using LangChain and ChatGPT 3.5 for human machine interface (HMI) alarming system for troubleshooting and predictive maintenance analysis. ChatGPTs understanding of Industry 5.0 may be basic and shallow. However, ChatGPT's answers are enlightening and helpful (Wang *et* al. [25]). Kiangala and Wang [15] didn't explicitly mention the hallucination of LLMs and the generation of contextual relevant response of LLM to the user's query in their work. They didn't develop a robust and stateful application using LangGraph for better control of their HMI system. Considering this gap, an attempt has been made in this paper to develop an economical ARAG-based system to generate contextual relevant response to the user's query in the light of I5.0 by using LangChain, LangGraph, and MISTRAL 7B. The proposed ARAG can be used for any other context to facilitate the interactions between man and machine. This paper gives an introduction to an adaptive retrieval-augmented generation (ARAG) method for text-based assistants for understanding of Industry 5.0 in section 1. In section 2, a novel approach for literature review is discussed lucidly. Section 3 discusses ARAG, a two-stage hybrid zero-resource hallucination detection approach, mathematical model of hallucination, template for binary evaluation of hallucination,

comparison between ARAG and RAG, multilingual ARAG, and architecture of ARAG. Section 4 gives the conclusion and future scope.

2. Literature Review

Literature review can be classified as narrative review, systematic review, scoping review, meta-analysis, integrative review, rapid review, critical review, content-centric review (Ghobakhloo *et* al. [11]), and multiple criteria decision analysis (MCDA)- based literature review. MCDA methods such as Analytic Hierarchy Process (AHP), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) etc. can be used to select relevant research papers (Steffen *et* al. [33]). In this paper, ScienceDirect, SCOPUS, Web of Science, and Google Scholar are used for searching for relevant research papers using keywords such as 'industry 5.0', 'industry 4.0', 'digital twin and industry 5.0', 'robot and industry 5.0', and 'blockchain and industry 5.0'. A total of 100 research papers written in English are initially collected, and 25 out of 100 are selected based on the content, citations, publishing journal's impact factor, the relevant work related to the use of advanced technology in the context of Industry 5.0 etc. Duplicate research papers are not considered. Table 1 shows some high-impact journals that were considered to prepare table 2. Initially, relevant papers are searched within the timeframe 2021 to 2024 using publish and perish software and important keywords are extracted from selected papers using VOSviewer, shown in fig. 1.

Table 1 Some high-impact journals that are considered for the literature review

Journal	Impact Factor
Journal of Innovation & Knowledge	15.6
IEEE/CAA Journal of Automatica Sinica	15.3
Journal of Manufacturing Systems	12.3
Technovation	11.1
Journal of Industrial Information Integration	10.4
IEEE Internet of Things Journal	8.2
Cleaner Logistics and Supply Chain	6.9
IEEE Open Journal of the Communications Society	6.3
Results in Engineering	6
Journal of King Saud University – Computer and Information Sciences	5.2

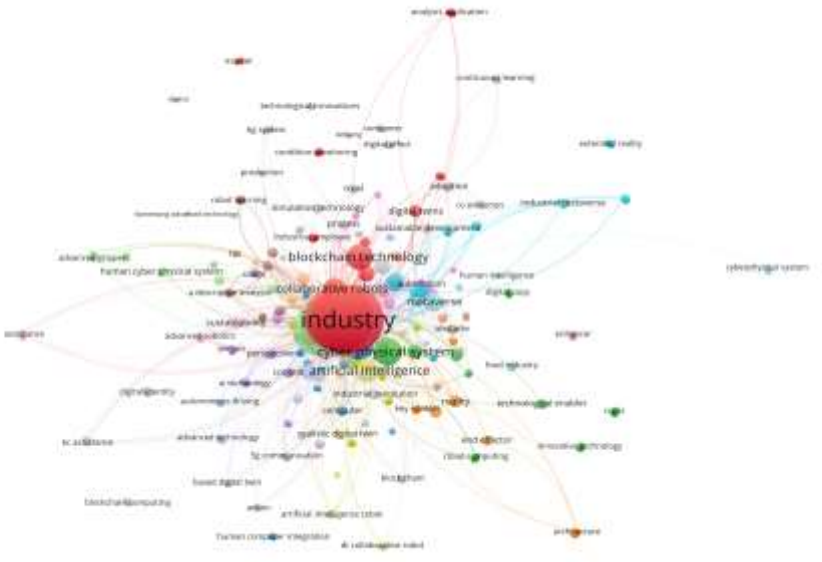

Fig 1 Some keywords in the light of Industry 5.0

Natural Language Processing (NLP) can take significant role in keyword extraction, contextual analysis, and summarization of relevant research papers and NLP could be used further for the literature review. In this paper, relevant research papers are searched and selected as per the following flow diagram, shown in fig 2:

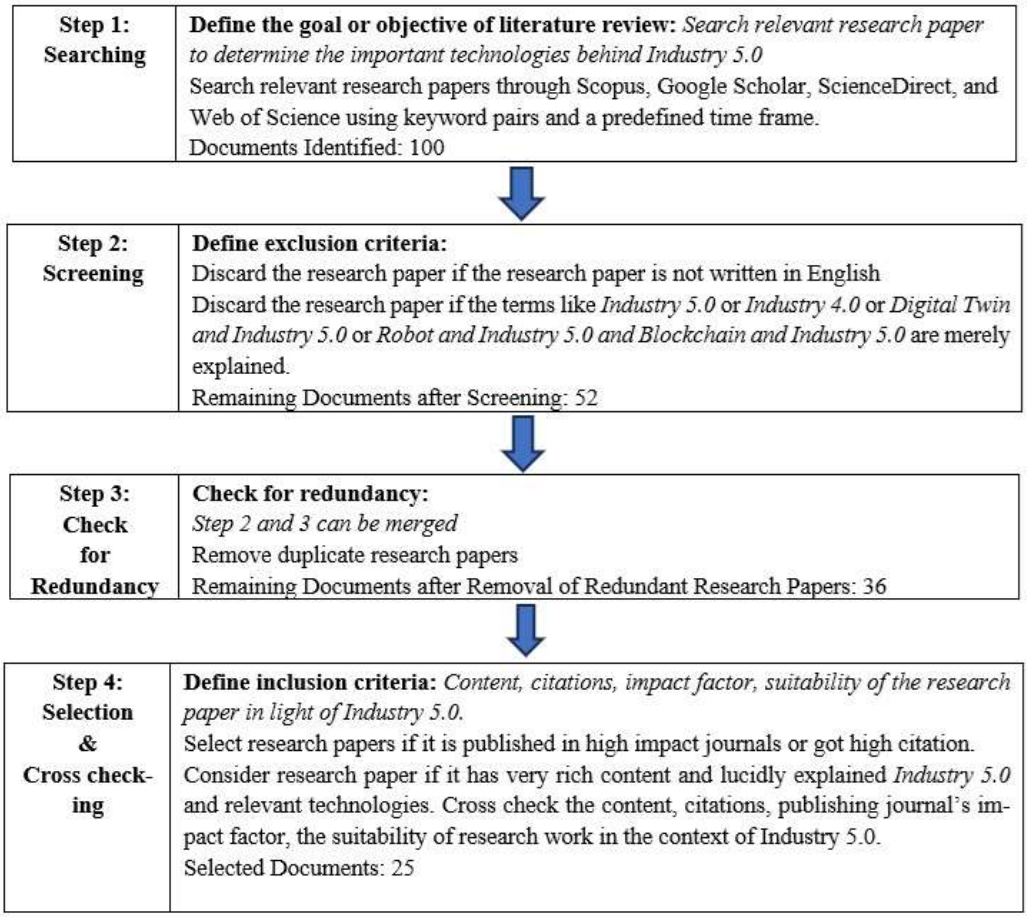

Fig 2 A pragmatic approach for literature review

Table 2 and fig 3 show the 22 most important technologies in the context of Industry 5.0. Fig. 1 gives a rough idea of some of the technologies in the context of Industry 5.0, whereas fig.3 gives a better interpretability or ranking of the most important technologies in the context of Industry 5.0. The proposed approach of literature review, thus, has significant benefits. Further, cross-checking has been used in the proposed approach during research paper selection to reduce decision bias, if any. Some of the technologies are discussed below in brief and interested readers can refer to relevant journals to have more details about the aforementioned 22 technologies.

2.1 Artificial Intelligence (AI): The definition of AI is based on two fundamental dimensions – 1. Human performance 2. Rationality [18] (p. 3-5). AI can be defined as a system that thinks and acts like humans by considering the human performance. Alternatively, AI can be defined as a system that thinks and acts rationally. The term rationally refers to doing the right things. In 1959, Arthur Samuel gave the definition of machine learning, a subset of AI, as follows: "Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed". Researchers have been trying to develop a replica of the human brain for long time. A Neuron or a nerve cell is the fundamental functional unit of the brain. A

neural network is the replica of neuron. Deep learning is a class of Machine Learning algorithms that uses neural network.

2.2 Internet of Things (IOT): It mainly controls processes by connecting devices, sensors such as Radio Frequency Identification (RFID) for real-time traceability, and uses technologies such as Near Field Communication Technology (NFC) for transmission and reception of short distance data, Wireless Sensor Networks (WSN) technology, and data storage (Yang *et* al. [29]).

2.3 Cloud Services and Mobile Edge Computing (MEC): Cloud services cater to the needs of different users by sharing computer hardware. IP addresses of computer hardware can't be detected on cloud. In I5.0, smart devices experience latency and battery challenges. In MEC servers, data processing and analysis are offloaded from smart devices to edge servers and thereby reduces latency and power consumption. MEC servers can also enhance the manufacturing system's performance, efficiency, real-time data processing analysis, decision-making etc. (Nauman *et* al. [16]).

2.4 Cyber-physical Systems (CPS): Cyber-physical system closely monitors real-time data to prevent data breaches, cyber-attack on physical systems, ransomware and malware attacks, eavesdropping etc. Safety and security have paramount importance during man and machine interactions in the collaborative environment (Alabdulatif *et* al. [2]).

2.5 Collaborative Robot (Cobot): These are specially designed robots to collaborate with human operators for providing support in various jobs (Alabdulatif *et* al. [2]).

Table 2 Some technologies and useful strategies for Industry 5.0

Authors	Year	Mass Customization & Waste Minimization	Digital Twin	DES	AI	XAI	ABS	ABM	AR/VR/ MR	CPS	IOT	IIOT	Solver	HMI / HCSM	Edge/ Cloud/ MEC	Cyber Security & Privacy	Cobot / Robotics	LLM	BCI	Big Data	BC	FL	GDM/ MCDA/MCDM
Turner and Garn	2022		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓						
Yang *et* al.	2024		✓		✓				✓		✓	✓		✓			✓		✓				
Nauman *et* al.[*1]	2024														✓		✓						
Barata and Kayser[*2]	2024		✓											✓			✓						
Ahmed *et* al.	2024			✓					✓	✓	✓					✓	✓	✓	✓	✓			
Xu *et* al.	2021				✓					✓				✓									
Shahbakhsha and Emada	2021				✓					✓	✓				✓					✓			
Zafar *et* al.[*3]	2023				✓						✓											✓	
Yang *et* al.[*4]	2024		✓		✓									✓									
Haghnazar *et* al.[*5]	2024	✓												✓									
Williamson and Prybutok[*6]	2024				✓															✓			
Garrido and Muniz	2024				✓				✓	✓	✓				✓		✓			✓			
Verma *et* al.	2022	✓	✓		✓					✓	✓				✓		✓			✓			
Alabdulatif *et* al.	2024				✓					✓	✓				✓	✓	✓			✓		✓	
Zhao *et* al.	2021				✓						✓				✓							✓	
Alonso *et* al.[*7]	2024				✓					✓	✓					✓							
Wang *et* al.	2023																						✓
Chand and Lu[*8]	2023													✓			✓						
Kiangala, and Wang	2024													✓				✓					
Sai *et* al.	2024																	✓					
Wang *et* al.	2023													✓				✓		✓			

❖ ABM: Agent Based Modelling, ABS: Agent Based Simulation, AI: Artificial Intelligence, XAI: Explainable AI, DES: Discrete Event Simulation, AR: Augmented Reality, VR: Virtual Reality, MR: Mixed reality, CPS: Cyber Physical Systems, Solver: Optimization Solver such as IBM ILOG Cplex, Gurobi, Open Source Solvers etc, IOT: Internet of Things, IIOT: Industrial Internet of Things, HMI: Human Machine Interface, HCSM: Human-Centric Smart Manufacturing, Edge: Edge Computing, MEC: Mobile Edge Computing, Cobot: Collaborative Robot, BCI: Brain Computer Interface, BC: Block Chain, FL: Federated Learning, *1: Research paper contains Mixed Integer Non-Linear Mathematical Model, *2: Research paper contains framework for digital twins in industry, *3 : Research paper contains application of Federated Learning (FL) and Convolution Gated Recurrent Unit (ConvGRU), *4 : Research paper contains the application of Finite Element Analysis (FEA) and ANSYS with detail discussion on different tools used for digital twin, *5: Research paper shows the use of digital fabrication in the mass customization of wood manufacturing, *6: Research paper integrates Neuromorphic Computing, Topological Data Analysis (TDA), Persistent Homology, Reinforcement Learning (RL) into sustainable supply chain, *7: Research paper contains RAMI 4.0, and three well-known definition of privacy, *8: Research paper proposed a mathematical model based on Learning-Forgetting-Fatigue- Recovery (LFFR) and solved it using NSGA-II.

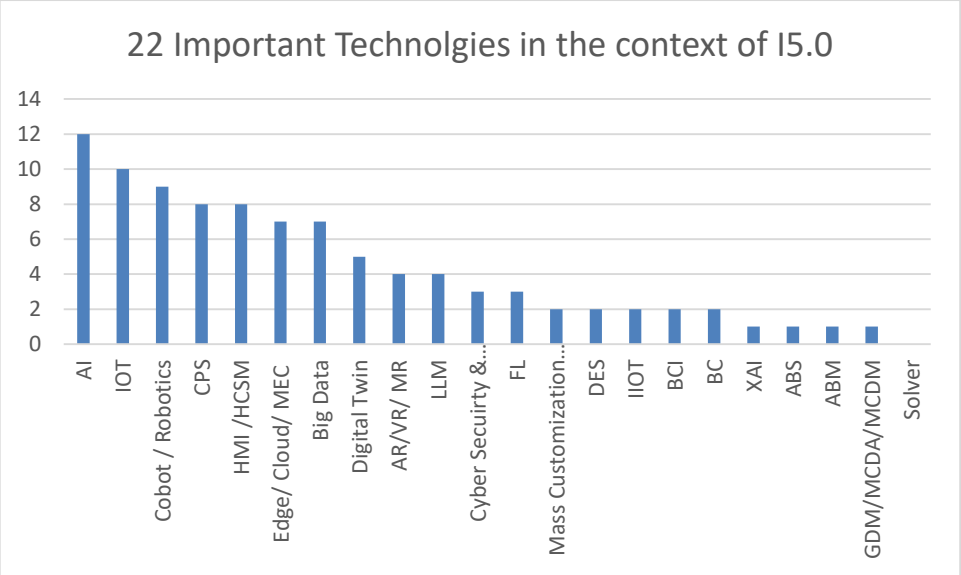

Fig 3 Important technologies in the context of Industry 5.0

The literature review shows that Artificial Intelligence (AI), Internet of Things (IOT), Collaborative Robot (Cobot), Cyber Physical System (CPS), Human Machine Interface (HMI), Edge Computing, Big Data, Digital Twin, Virtual Reality (VR), Large Language Model etc. are the most important technologies in the context of I5.0. The study reveals that Multiple Criteria Decision Analysis (MCDA), Multiple Criteria-Based Decision Method (MCDM), and Group Decision Method (GDM) are also important for having consensus on using a specific technology for I5.0. The study further reveals that I5.0 demands predictive as well as prescriptive analysis. Predictive analysis uses AI for forecasting. Prescriptive analysis, on the other hand, uses optimization solvers with predictive analysis. It is pertinent to mention that I5.0 is expected to use multiple criteria-based optimization methods with AI to enhance the quality, safety, and security of a man-machine collaborative environment. In the human-machine interface, chatbots, voice-based assistants, and text-assistants play the utmost important role in giving better interactions between man and machine. The Large Language Model (LLM), a transformer-based AI model, has become the backbone of text-assistants since its inception. Finally, Reinforcement Learning (RL), a subset of AI, and LLM will significantly influence the man-machine collaborative environment.

3. Adaptive RAG

"*Adaptive Retrieval-Augmented Generation (RAG) can offer a robust middle ground among the iterative LLM for augmentation method for complex queries, single-step method for simpler queries, and even no-retrieval-augmented method for the most straight-forward queries*" (Jeong *et* al. [13])

If **q** is the query and **a** is the response of LLM then the response can be expressed as follows:

a = LLM(**q**). ……. (1)

This is known as no-retrieval-augmented method. If the E is the external knowledge source and $\boldsymbol{d}$ is the retrieved documents then **d** can be expressed as follows:
d = Retriever (**q,** E) where **d** ∈ E…… (2)

This is known as simple RAG method or single-step approach (Jeong *et* al. [13]). Usually, single-step

approach demands considerable collection of documents to answer query effectively. Adaptive Retrieval-Augmented Generation (ARAG) is more effective if answers of some queries are not be available in the external knowledge source.

3.1 Hallucination

The term "hallucination" refers to the response of a LLM that contains plausible but factually incorrect or nonsensical information, and it is inevitable and an innate limitation of LLMs regardless of model architecture, prompting techniques, training data, and learning algorithms (Xu *et* al. [27]).

According to Xu *et* al. [27] if S be a computable set of all the finite-length strings and f is the ground truth function for a given input string s where f(s) is the correct output ∀ s ∈ S. Hallucination can be mathematically defined as follows:

$\exists s \in S \; such \; that \;\; h(s) \neq f(s) \;\; where \; h(s)$ is the response generated by the LLM …… (3)

3.2 A Two-Stage Factual Incorrectness or Nonsensical Information Detection Approach:

Hallucination can be determined by using Black-Box Uncertainty Quantification (UC) approach (Manakul *et* al. [37]), White-Box UC approach (Manakul *et* al. [37], Azaria *et* al. [38]), LLM-as-a-Judge (Bai *et* al. [39], Hofsatter *et* al. [40]), Ensemble approach (Bouchard and Chauhan [41]) etc.

In this paper, a two-stage hybrid zero-resource hallucination detection approach has been developed for the detection of factual incorrectness or nonsensical information in the generated response. Initially, a binary classifier detects the factual incorrectness in the generated response w.r.t the relevant retrieved documents of RAG. A binary classifier has been used to grade the retrieved documents. If the classifier rejects the retrieved documents, then instructions are given to the system to retrieve the documents again. In the second stage, a panel of LLM-judges are used to cross-examine the generated response w.r.t the question and LLMs don't access internet, database of source content, or ground truth texts during cross-examination and thus, the second-stage, is a zero-resource hallucination detection approach (Bouchard and Chauhan [41]). The responses of binary classifier and the responses of LLM-judges are compared with the responses of human annotators.

3.2.1 Prompt Based Binary Evaluation of Hallucination

Prompt based factuality evaluation such as pairwise comparison, Likert scale soring, binary evaluation etc. are quite effective (Gao *et* al. [10]). Please refer to the prompt for binary evaluation of hallucination, shown in fig 4.

```
You are an expert in determining hallucination in the summary as given below.
    Summary = {summary}
    Documents = {documents}
    Is there any hallucination in the Summary with respect to the Documents?
    Write no statement, no additional_kwargs, and no response_metadata in your answer.
    'Yes' means there is hallucination.
    'No' means there is no hallucination.
    Your answer will be in string datatype format.
    Only write content = 'Yes' or 'No'.
```

Fig 4. The template for binary evaluation of hallucination

3.2.2 A Panel of LLM Judges for Evaluation of Hallucination

A single LLM can have decision bias. A diverse collection of LLMs, on the other hand, can overcome the judgmental bias. The selection of LLM-judges is an open issue. In this paper, Mistral 7B, Llama3 8B, Llama2 7B, and Deepseek-R1 are considered. Mistral 7B and Llama3 8B outperformed Deepseek-R1. Each LLM-judge cross-examined the generated response w.r.t the user's query and gave a score using linguistic variables, shown in table 3a. Triangular Fuzzy Number (TFN) can also be used to represent linguistic variables. Table 3b shows the LLMs that are considered for generating responses and detecting hallucination in the generated response. A voting function, VF, is used give aggregate score. It is pertinent to mention that in Multiple Criteria Decision Analysis (MCDA), geometric mean is usually considered to overcome decision bias. The mathematical expression of VF is shown below:

$L_{Judge} : Set\ of\ LLMs$ ……(4)
$Priority_i : Priority\ given\ to\ i^{th}\ LLM\ judge$ ……(5)
$Score_i : Score\ given\ by\ i^{th}\ LLM\ judge\ in\ 5\ point\ scale$ ……(6)
$Normalize\_Score_i : Normalize\ score\ of\ the\ i^{th}\ LLM\ Judge$ ……(7)

$$VF = \left(\sum_{i \in L_{Judge}} Priority_i \times Normalize\_Score_i\right)^{1/no.of\ judges} \quad \text{……(8)}$$

$$where\ Normalize\_Score_i = \frac{Score_i}{\sum_{i \in L_{Judge}} Score_i}$$

Table 3a The 5-Point Agreement Scale

Linguistic Variables	Numeric Value
Strongly Disagree	1
Disagree	2
Neutral	3
Agree	4
Strongly Agree	5

Table 3b Response Generator and Judge Models

Response Generator	**Mistral 7B used for ARAG**
Prompt-Based Binary Classifier	Mistral 7B
LLM Panel of Judges	Mistral 7B, Llama3 8B, Llama2 7B, and Deepseek R1

This study further shows the alignment of LLM-judges with human annotators, shown in fig 5. In fig 5, error bar refers to the standard deviation of the score achieved by different LLM models.

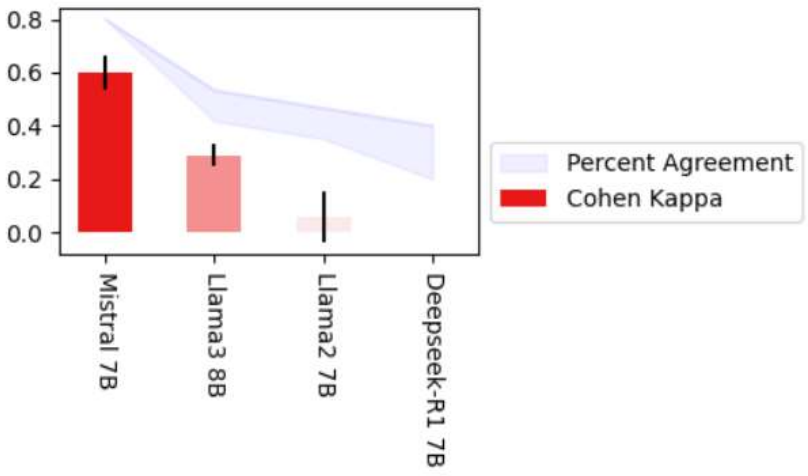

Fig 5 The average score achieved by the LLM-judges and their alignment with human annotators

Table 4 Correlation between LLM judges and Human Judgments

Judge	Kendal Tau	Pearson
Mistral 7B	0.744	0.80
Llama3 8B	0.574	0.602
Llama2 7B	0.378	0.375

A strong positive correlation has been shown by Mistral 7B with human judgments, shown in Table 4. Deepseek -r1 shows very poor correlation with human judgements and, thus, discarded from the panel of LLM-judges.

3.2 Architecture of ARAG

An alternative approach would be to use the artificial intelligence-based search engine and RAG simultaneously. A binary classifier will determine the use of an artificial intelligence-based search engine or RAG for a given query. The crux of ARAG is a binary classifier which will determine whether the system will retrieve the response to a user's query from the artificial intelligence-based search engine or retrieve the relevant documents from the vector store database. This binary classifier is prompt-based. A prompt is basically a set of instructions to LLM. In the proposed ARAG, multiple binary classifiers are used to verify the relevance of the retrieved documents, relevance of the response generated by LLM, reformation of the user's query to get a better response from the LLM, and also for testing of hallucination. If C is a binary classifier, then

$$reponse = \begin{cases} 1, & use\ RAG \\ 0, & use\ search\ engine \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

response = C (**q**)……………………………………………………… (10)

d = Retriever (**q,** E) where **d** ∈ E if response = 1………… (11)

d = Web_Search(**q**) if response = 0……………. (12)

In this paper, MISTRAL 7B LLM, an advanced AI model specialized in Natural Language Processing (NLP), is considered instead of ChatGPT as it is free. Here, 7B refers to the seven billion parameters. The proposed ARAG could use an artificial intelligence-based internet search engine to give the best result to the user's query, if RAG fails to generate relevant answers. LLM developers generally split large pieces of input text into smaller segments. This is known as chunking. A computer converts text to a n-dimensional numeric value or vector for similarity search. LLM embeddings are basically the vector representation of words, phrases, or the entire text. In the conversational interface, the system should quickly retrieve relevant documents for the user's query. This is achieved by using indexing. Chunking, indexing, and retrieval of large collections of documents are costly and memory-sensitive. Ollama embeddings, FAISS, and MISTRAL 7B have been used considering the cost and memory constraints. Facebook AI Similarity Search (FAISS) is a vector database which is commonly used for fast retrieval of documents from the vector store or vector database. A graph-based model is further developed for better control of the ARAG, shown in fig 8. LangGraph is an open-source framework for building complex, stateful, multi-agent applications using LLMs. LangGraph agents can do linguistic tasks effectively with human beings, and it is extremely important for conversational interface to give a seamless collaboration between man and machine. The LangGraph agent brings more flexibility by determining the best strategy for effective action by using specialized tools. The author strongly suggests using LangGraph for the development of a complex conversational interface between man and machine. This conversational interface could be used for a human-centric

manufacturing process. Finally, Mistral 7B, FAISS, panel-of-judges, asynchronous processing, caching etc are used for low latency.

3.2.2 ARAG vs RAG

RAG may fail to generate a relevant answer to the user's query. It may generate an absurd response due to hallucinations. The response of RAG mainly depends on the corpora. ARAG uses internet search along with RAG. If RAG generates an irrelevant response, then ARAG will use the internet for a relevant answer. If an internet search generates any irrelevant response, then the binary classifier will cross-check and search the internet again, shown in fig. 8.

3.2.3 Multilingual ARAG

The detail research on multilingual responses is out of scope of this paper. However, a brief introduction has been given to show the capability of Mistral 7B to generate response in multiple languages. Multilingual large language models (MLLMs) use advanced large language models to respond to user's queries across multiple languages (Qin *et* al. [34]). Mistral is a multilingual large language model and has achieved significant success in the polyglot task. A finetuned Mistral outperforms ChatGPT "gpt-3.5-turbo" in zero shot translation (Moslem *et* al. [35]). Ono and Moirta [36] conducted a study to show the capabilities of large language models such as ChatGPT-4, Google Gemini, and Mistral and confirmed the excellent proficiency of CharGPT-4 in English, Google Gemini outperformed ChatGPT-4 in Japanese, and Mistral 7B showed a balanced performance in English and Japanese. Fig. 6 shows the response of basic Mistral 7B in Bengali and French.

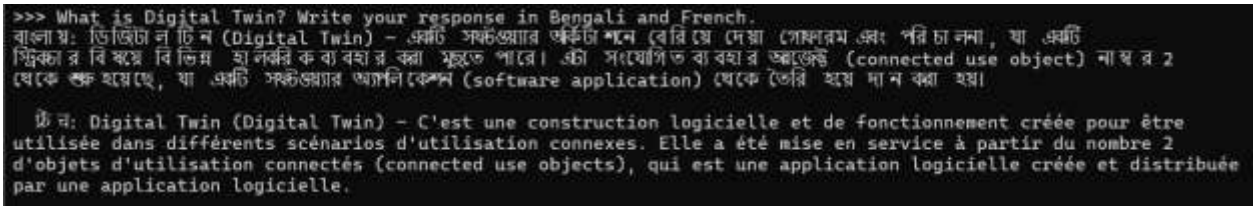

Fig 6 The performance of basic Mistral 7B in Bengali and French

In this regard, the high accuracy of the basic Mistral 7B can be achieved by finetuning the model. Moslem *et* al. [35] used adaptive machine translation to enhance the performance of basic Mistral 7B in Spanish-to-English translation. The proposed ARAG shows excellent performance in the polyglot task and gives context-relevant response compared to the basic Mistral 7B in light of Industry 5.0, shown in fig. 7.

Fig 7 Response of ARAG in French

Further, a finetuned model of Mistral 7B has been developed to compare the performance with ARAG. A P100 GPU has been used for a small dataset in this regard**. A finetuned model can generate contextually relevant responses in light of Industry 5.0 but needs more computational resources and, thus, can't be considered as economical as ARAG. A small dataset containing questions and answers in English and Hindi has been used in this regard. This study shows that finetuned Mistal 7B model can generate contextually relevant responses in English but fails to interpret questions in Hindi.

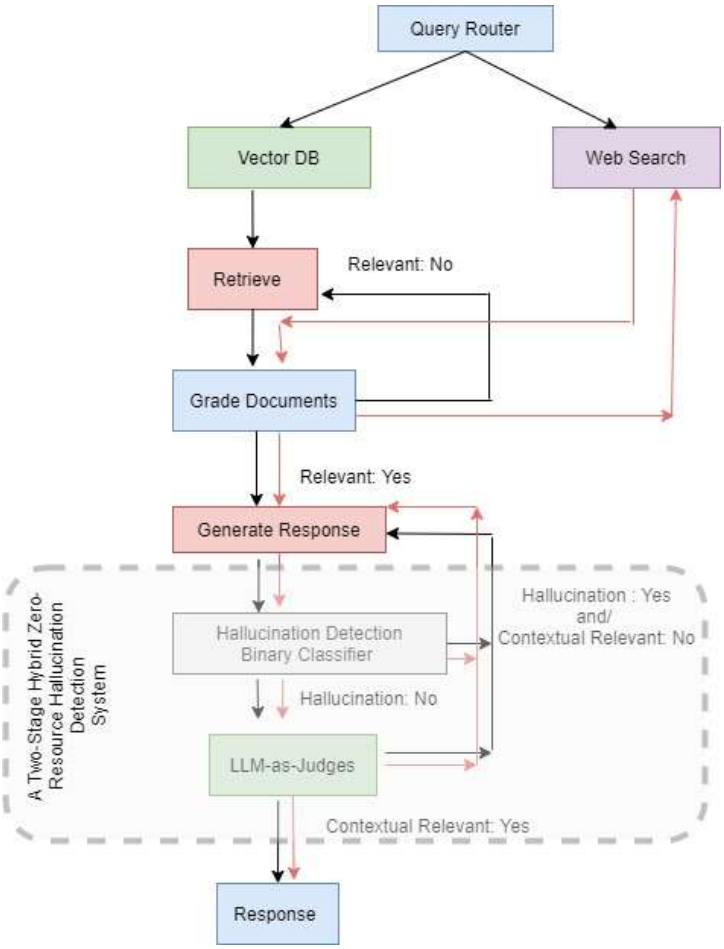

Fig 8 Schematic diagram of the proposed ARAG

The proposed ARAG can respond to simple and complex queries. It can automatically change the search mode from 'vector store' to 'web search' to give a context-relevant response to the user's query, shown in fig. 9 and 10.

** https://www.kaggle.com/code/krishm/mistral-lora-industry5?scriptVersionId=242788185

Fig 9 Response of ARAG to simple query

Fig 10 Response of ARAG to complex query

4 Conclusion and Future Scope

This study reveals that Mistral 7B-based ARAG is an excellent multilingual and economical question-answering system. The proposed multilingual ARAG can generate human-like context-relevant responses in English, French etc. A finetune model has also been developed to compare the performance with ARAG. Finetuned model can give better security to the corpora of RAG, but it needs more computational resources thus shouldn't be used to develop an economical system. The latency is also expected to be reduced further. Further study is needed to tradeoff latency and cost, and it is considered as the future scope of the work. The proposed system can be used for transportation, hospitality, supply chain and other sectors as a text-based human-machine interface.

The key findings of the paper are as follows:

- Mistral 7B is an excellent multilingual human-like text generator.
- This paper shows a novel approach for selecting research papers and further shows that Artificial Intelligence (AI), Internet of Things (IOT), Collaborative Robot (Cobot), Cyber Physical System (CPS), Human Machine Interface (HMI), Edge Computing, Big Data, Digital Twin, Virtual Reality (VR) etc. are the most important technologies in the context of I5.0. Reinforcement Learning (RL), a subset of AI, and LLM is going to play the most important role in the context of I5.0.

- This paper shows a two-stage zero-resource hallucination detection approach by integrating a binary classifier and a panel of LLM judges. Mathematical expressions are also included for the better understanding of hallucination detection approach.
- The proposed ARAG can generate contextually relevant human-like text as per the user's query in English, French etc.

References

1. Ahmed I, Hossain NUI, Fazio SA, Lezzi M, Islam MdS. A decision support model for assessing and prioritization of industry 5.0 cybersecurity challenges. Sustainable Manufacturing and Service Economics [Internet]. 2024 Jan 1;3:100018. Available from: https://www.sciencedirect.com/science/article/pii/S266734442400001X
2. Abdullah Alabdulatif, Navod Neranjan Thilakarathne, Zaharaddeen Karami Lawal. A Review on Security and Privacy Issues Pertaining to Cyber-Physical Systems in the Industry 5.0 Era. Computers, materials & continua/Computers, materials & continua (Print) [Internet]. 2024 Jan 1;80(3):3917–43. Available from: https://www.sciencedirect.com/org/science/article/pii/S1546221824006507
3. Alonso R, Haber RE, Castaño F, Diego Reforgiato Recupero. Interoperable Software Platforms for Introducing Artificial Intelligence Components in Manufacturing: A Meta-Framework for Security and Privacy. Heliyon. 2024 Feb 1;10(4):e26446–6.
4. Ansari F, Erol S, Sihn W. Rethinking Human-Machine Learning in Industry 4.0: How Does the Paradigm Shift Treat the Role of Human Learning? Procedia Manufacturing. 2018;23:117–22.
5. Barata J, Kayser I. How will the digital twin shape the future of industry 5.0? Technovation. 2024 Jun 1;134:103025–5.
6. Breque M, De Nul L, Petridis A. Industry 5.0 Towards a sustainable, Humancentric and Resilient European Industry [Internet]. European Commission; 2021 [cited 2025 Jun 3]. Available from: https://eurocid.mne.gov.pt/sites/default/files/repository/paragraph/documents/17991/brochura-industry-50_0.pdf
7. Chang Y, Wang X, Wang J, Yuan W, Yang L, Zhu K, et al. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology. 2024 Jan 23;15(3):1–45.
8. Chand S, Lu Y. Dual task scheduling strategy for personalized multi-objective optimization of cycle time and fatigue in human-robot collaboration. Manufacturing Letters. 2023 Aug;35:88–95.
9. Garrido S, Muniz J, Batista Ribeiro V. Operations Management, Sustainability & Industry 5.0: A critical analysis and future agenda. Cleaner Logistics and Supply Chain [Internet]. 2024 Mar 1;10:100141. Available from: https://www.sciencedirect.com/science/article/pii/S2772390924000039
10. Gao M, Ruan J, Sun R, Yin X, Yang S, Wan X. Human-like Summarization Evaluation with ChatGPT [Internet]. 2023. Available from: https://arxiv.org/pdf/2304.02554
11. Morteza Ghobakhloo, Iranmanesh M, Tseng ML, Andrius Grybauskas, Stefanini A, Azlan Amran. Behind the definition of Industry 5.0: a systematic review of technologies, principles, components, and values. Journal of Industrial and Production Engineering. 2023 May 27;40(6):1–16.
12. Ramtin Haghnazar, Yasaman Ashjazadeh, Hauptman J, Nasir V. A Computational Design Integrated Digital Fabrication Framework for Mass Customization in Industry 5.0 Manufacturing with Non-Standard Natural Materials. Results in Engineering. 2024 Jun 11;23:102400–0.
13. Jeong S, Baek J, Cho S, Hwang SJ, Park J. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. arXiv (Cornell University) [Internet]. 2024 Jan 1 [cited 2025 Jun 3]; Available from: https://aclanthology.org/2024.naacl-long.389/
14. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? on Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences [Internet]. 2023 Apr 1;103. Available from: https://www.sciencedirect.com/science/article/abs/pii/S1041608023000195
15. Kahiomba Sonia Kiangala, Wang Z. An experimental hybrid customized AI and generative AI chatbot human machine interface to improve a factory troubleshooting downtime in the context of Industry 5.0. The International Journal of Advanced Manufacturing Technology. 2024 Apr 3;132.

16. Nauman A, Khan WU, Ghadah Aldehim, Alqahtani H, Nuha Alruwais, Mesfer Al Duhayyim, et al. Communication and computational resource optimization for Industry 5.0 smart devices empowered by MEC. Journal of King Saud University - Computer and Information Sciences. 2023 Dec 15;36(1):101870–0.
17. Rožanec JM, Novalija I, Zajec P, Kenda K, Tavakoli Ghinani H, Suh S, et al. Human-centric artificial intelligence architecture for industry 5.0 applications. International Journal of Production Research. 2022 Nov 7;61(20):1–26.
18. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. Prentice Hall, New Jersey, ISBN 0-13-103805-2.; 2003.
19. Sai S, Sai R, Vinay Chamola. Generative AI for Industry 5.0: Analyzing the impact of ChatGPT, DALLE, and Other Models. IEEE open journal of the Communications Society. 2024 Jan 1;1–1.
20. Shahbakhsh M, Emad GR, Cahoon S. Industrial revolutions and transition of the maritime industry: The case of Seafarer's role in autonomous shipping. The Asian Journal of Shipping and Logistics. 2022 Mar;38(1):10–8.
21. Turner CJ, Garn W. Next generation DES simulation: A research agenda for human centric manufacturing systems. Journal of Industrial Information Integration. 2022 Jul;28:100354.
22. Verma A, Bhattacharya P, Madhani N, Trivedi C, Bhushan B, Tanwar S, et al. Blockchain for Industry 5.0: Vision, Opportunities, Key Enablers, and Future Directions. IEEE Access. 2022;10:69160–99.
23. Ljubo Vlacic, Huang H, Mariagrazia Dotoli, Wang Y, Ioannou PA, Fan L, et al. Automation 5.0: The Key to Systems Intelligence and Industry 5.0. IEEE/CAA Journal of Automatica Sinica. 2024 Jul 19;11(8):1723–7.
24. Wang ZJ, Chen Z, Xiao L, Su Q, Govindan K, Skibniewski MJ. Blockchain adoption in sustainable supply chains for Industry 5.0: A multistakeholder perspective. Journal of Innovation & Knowledge. 2023 Oct 1;8(4):100425–5.
25. Wang FY, Yang J, Wang X, Li J, Han QL. Chat with ChatGPT on Industry 5.0: Learning and Decision-Making for Intelligent Industries. IEEE/CAA Journal of Automatica Sinica [Internet]. 2023 Apr 1;10(4):831–4. Available from: https://ieeexplore.ieee.org/abstract/document/10085975/
26. Williamson SM, Prybutok V. Integrating human-centric automation and sustainability through the NAToRM framework: A neuromorphic computing approach for resilient industry 5.0 supply chains. International Journal of Information Management Data Insights. 2024 Nov 1;4(2):100278–8.
27. Xu Z, Jain S, Kankanhalli M. Hallucination is Inevitable: An Innate Limitation of Large Language Models [Internet]. 2024. Available from: https://arxiv.org/pdf/2401.11817
28. Xu X, Lu Y, Vogel-Heuser B, Wang L. Industry 4.0 and Industry 5.0—Inception, conception and perception. Journal of Manufacturing Systems [Internet]. 2021 Oct;61(1):530–5. Available from: https://www.sciencedirect.com/science/article/pii/S0278612521002119
29. Yang J, Liu Y, Morgan PL. Human–machine interaction towards Industry 5.0: Human-centric smart manufacturing. Digital Engineering. 2024 Aug 1;2:100013–3.
30. Yang T, Razzaq L, H. Fayaz, Qazi A. Redefining fan manufacturing: Unveiling industry 5.0's human-centric evolution and digital twin revolution. Heliyon. 2024 Jul 1;10(13):e33551–1.
31. Zafar MH, Bukhari SMS, Abou Houran M, Moosavi SKR, Mansoor M, Al-Tawalbeh N, et al. Step towards secure and reliable smart grids in Industry 5.0: A federated learning assisted hybrid deep learning model for electricity theft detection using smart meters. Energy Reports [Internet]. 2023 Nov 1;10:3001–19. Available from: https://www.sciencedirect.com/science/article/pii/S2352484723013458
32. Zhao Y, Zhao J, Jiang L, Tan R, Niyato D, Li Z, et al. Privacy-Preserving Blockchain-Based Federated Learning for IoT Devices. IEEE Internet of Things Journal. 2020;8(3):1817–29.
33. Steffen V, de Oliveira MS, Trojan F. A Novel Approach for Systematic Literature Reviews Using Multi-Criteria Decision Analysis. Journal of Intelligent Management Decision. 2024 May 23;3(2):116–38.
34. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, et al. A survey of multilingual large language models. Patterns. 2025 Jan;6(1):101118.
35. Moslem Y, Haque R, Way A. Fine-tuning Large Language Models for Adaptive Machine Translation. arXiv (Cornell University). 2023 Dec 19;
36. Ono K, Morita A. Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked Against MMLU. 2024 Mar 4 [cited 2024 Nov 2]; Available from:

https://www.techrxiv.org/users/748222/articles/719880-evaluating-large-language-models-chatgpt-4-mistral-8x7b-and-google-gemini-benchmarked-against-mmlu
[37] Manakul P, Liusie A, Gales MJF. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:230308896 [cs] [Internet]. 2023 Mar 15; Available from: https://arxiv.org/abs/2303.08896
[38] Azaria A, Mitchell T. The Internal State of an LLM Knows When its Lying [Internet]. arXiv.org. 2023. Available from: https://arxiv.org/abs/2304.13734
[39] Bai Y, Ying J, Cao Y, Xin Lv, He Y, Wang X, Yu J, Zeng K, Xiao Y, Lyu H, Zhang J, Li J, and Hou L. Benchmarking Foundation Models with Language-Model-as-an-Examiner. arXiv (Cornell University). 2023 Jun 7
[40] Verga P, Hofstatter S, Althammer S, Su Y, Piktus A, Arkhangorodsky A, Xu M, and Lewis N W P. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models [Internet]. arXiv.org. 2024. Available from: https://arxiv.org/abs/2404.18796
[41] Bouchard D and Chauhan MS. Uncertainty Quantification for Language Models: A Suite of Black-Box, White-Box, LLM Judge, and Ensemble Scorers [Internet]. arXiv.org. 2025 [cited 2025 Jul 4]. Available from: https://arxiv.org/abs/2504.19254