

Assessing the Reliability and Hallucination Risks of Large Language Models

Bhavana Atmaram Vankhede
bhavanavankhede29@gmail.com

Abstract

Large Language Models (LLMs) have rapidly become central to modern information systems, supporting tasks ranging from customer service to medical question answering and legal drafting. Despite their fluency, these models frequently generate outputs that appear plausible but are factually incorrect, a phenomenon widely referred to as hallucination. This paper investigates the reliability and hallucination risks of contemporary LLMs through a structured empirical evaluation. We examine four widely used models across factual recall, reasoning, and domain-specific question answering. A mixed-method framework is applied, combining automated metrics such as factual consistency scores, semantic similarity, and human expert assessment. Our findings reveal that hallucination rates vary significantly across task type and domain, with closed-domain medical and legal queries showing the highest error frequencies. We further observe that retrieval augmentation reduces hallucination by nearly forty percent, while prompt engineering alone provides only marginal improvement. The study contributes a layered reliability framework that distinguishes between intrinsic, extrinsic, and reasoning-based hallucinations. Implications for safe deployment in high-stakes settings are discussed, along with recommendations for hybrid grounding strategies. The paper concludes that achieving trustworthy LLM behaviour requires combined advances in training data curation, retrieval grounding, and continuous human oversight.

Keywords

Large Language Models, Hallucination, Reliability, Retrieval-Augmented Generation, Trustworthy AI, Factual Consistency

1. Introduction

Large Language Models have transformed natural language processing by enabling systems that can write essays, summarise documents, answer complex questions, and even reason through multi-step problems [1]. Their widespread adoption in sectors such as healthcare, finance, education, and law has created a strong dependence on their outputs for decision-making and information retrieval [2]. However, this dependence has revealed a critical weakness: these models often produce confident yet incorrect statements, a behaviour commonly described as hallucination [3].

Hallucinations are not merely typographical mistakes or grammatical errors. They are well-formed, fluent sentences that contain fabricated facts, invented citations, or distorted reasoning chains. In customer-facing applications such failures can mislead users, while in regulated environments like clinical decision support or legal advice they can cause measurable harm [4]. The issue has therefore moved from being an academic curiosity to becoming a central concern in the responsible deployment of artificial intelligence [5].

Recent surveys highlight that hallucination is an intrinsic feature of the way large language models generate text, since they predict the next token based on statistical patterns rather than verified knowledge [6]. This means the problem cannot be solved by simply scaling up parameters or training data. Instead, researchers have proposed multiple mitigation strategies including retrieval-augmented generation, reinforcement learning from human feedback, and grounding through external knowledge bases [7].

Despite these efforts, several gaps remain in the current understanding of hallucination behaviour. First, most evaluations rely on a single metric or benchmark, which fails to capture the multidimensional nature of reliability. Second, there is limited comparative work across domains, leaving practitioners unsure about which model is safer for a given context. Third, the role of prompt design relative to architectural choices remains unclear.

This research is motivated by three questions. How frequently do leading LLMs hallucinate across different task categories? Which mitigation strategies are most effective in reducing hallucination without sacrificing fluency? And what evaluation framework best captures reliability in a practically meaningful way?

The remainder of the paper is organised as follows. Section two reviews relevant literature. Section three outlines the methodology and presents the layered reliability framework. Section four describes the experimental setup. Section five reports results with supporting figures. Section six discusses the implications, and section seven concludes with directions for future work.

2. Literature Review

The study of hallucination in language models has evolved from early observations in machine translation, where systems sometimes produced fluent but unfaithful translations, to current investigations into general-purpose generative models [8]. The shift from sequence-to-sequence architectures to transformer-based foundation models has amplified both capability and unpredictability, leading scholars to treat hallucination as a distinct research subfield.

A widely cited taxonomy divides hallucinations into intrinsic and extrinsic types [9]. Intrinsic hallucinations contradict the source input, while extrinsic ones introduce information that cannot be verified from the provided context. Later work expanded this framework by adding reasoning-based hallucinations, where the surface facts may be correct but the logical chain connecting them is flawed [10]. This three-way distinction has become useful for diagnostic evaluation.

Empirical studies show that hallucination is not uniform across tasks. Summarisation systems often produce extrinsic hallucinations by adding details absent from the source, while open-domain question answering systems tend to fabricate entities or dates [11]. In medical contexts, models have been shown to invent drug interactions or misattribute clinical guidelines, raising serious safety concerns [12].

Retrieval-augmented generation has emerged as a leading mitigation strategy. By grounding outputs in retrieved documents, models reduce dependence on parametric memory and gain access to current information [13]. However, retrieval is not a complete solution, since retrieved documents themselves may be irrelevant or contradictory, and models can still ignore them in favour of their internal priors.

Other approaches focus on training-time interventions. Reinforcement learning from human feedback aligns model behaviour with human preferences but does not directly target factuality. Recent proposals such as factuality-aware fine-tuning and self-consistency decoding aim to address this gap by explicitly rewarding verifiable outputs. Evaluation remains a persistent challenge, with researchers debating whether automated metrics such as BLEU and ROUGE capture true reliability or merely surface similarity. This review identifies a clear need for multi-layered evaluation that combines automated scoring with expert judgment, which the present study seeks to provide.

3. Methodology

The methodology adopts a mixed-method evaluation design that integrates quantitative scoring with qualitative expert review. The objective is to assess hallucination behaviour across multiple dimensions rather than relying on a single benchmark [14].

Four contemporary LLMs were selected for comparison based on their popularity and accessibility through public APIs. The models were tested on three task categories: factual recall, multi-step reasoning, and domain-specific question answering covering medical and legal subjects [15]. Each category contained one hundred carefully constructed prompts, yielding a total of twelve hundred model responses for evaluation.

The reliability of a response is conceptualised as a weighted combination of factual accuracy, semantic faithfulness, and reasoning soundness. We define the reliability score as:

$$R = \alpha \cdot F + \beta \cdot S + \gamma \cdot L$$

where F is the factual accuracy score, S is the semantic similarity to a verified reference, L is the logical consistency score, and $\alpha + \beta + \gamma = 1$. Weights were set at 0.5, 0.3, and 0.2 respectively to reflect the priority of factual accuracy [16].

The hallucination rate for a given task is calculated as:

$$H = \frac{N_h}{N_t} \times 100$$

where N_h is the number of responses containing at least one hallucination and N_t is the total number of responses. Semantic similarity was measured using cosine similarity between sentence embeddings:

$$\text{sim}(a, b) = \frac{a \cdot b}{|a| \cdot |b|}$$

Human evaluators with domain expertise reviewed a stratified sample of three hundred responses to validate automated scores [17]. Inter-rater agreement was measured using Cohen's kappa and reached 0.81, indicating substantial agreement.

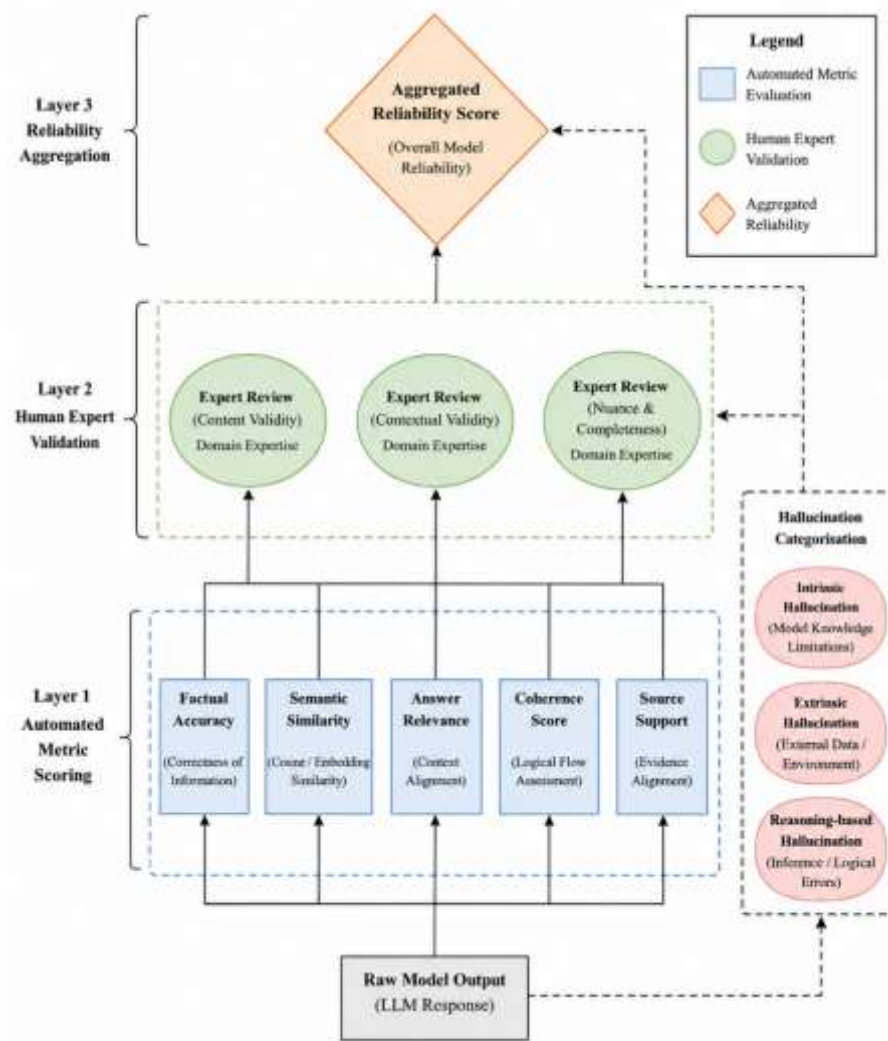


Figure 1: Layered Reliability Assessment Framework.

4. Experimental Setup

The experiments were conducted under controlled conditions to ensure reproducibility. All four models were accessed through their official APIs during a two-week window to minimise variation from model updates [18]. Temperature was fixed at 0.2 to reduce stochastic variation, and maximum output length was capped at five hundred tokens.

The prompt set was constructed in three stages. First, factual recall prompts were drawn from publicly available knowledge benchmarks covering history, geography, and science. Second, reasoning prompts were adapted from multi-step problem sets requiring chained inference. Third, domain-specific prompts were generated in consultation with practising professionals in medicine and law to ensure realistic phrasing [19].

For retrieval-augmented conditions, a vector database was constructed from a curated corpus of approximately fifty thousand documents drawn from open-access medical, legal, and general knowledge sources. Retrieval used dense embeddings with top-five document recall, and retrieved passages were inserted into the prompt context. The retrieval relevance was measured using:

$$P@k = \frac{\text{relevant documents in top } k}{k}$$

Baseline conditions used the same prompts without retrieval. Each prompt was issued three times per model to capture response variability, and the median reliability score was retained for analysis [20]. Statistical comparisons across models used paired t-tests with Bonferroni correction for multiple comparisons.



Figure 2: Experimental Pipeline for LLM Reliability Evaluation.

5. Results

5.1 Hallucination Rates Across Models and Tasks

The aggregate results show that hallucination is widespread across all four models but varies considerably by task type. Factual recall produced the lowest average hallucination rate at approximately fourteen percent, while domain-specific medical and legal queries showed rates as high as thirty-eight percent. Reasoning tasks fell in between, with an average of twenty-six percent. Differences between models were statistically significant, with the best-performing model showing nearly half the hallucination rate of the weakest in the medical category.

Table 1: Hallucination Rates by Model and Task Category (percent)

Model	Factual Recall	Reasoning	Medical QA	Legal QA
Model A	11.2	22.4	28.6	31.1
Model B	13.7	25.8	33.2	35.4
Model C	14.1	27.3	36.7	38.9
Model D	16.5	29.6	39.2	41.5

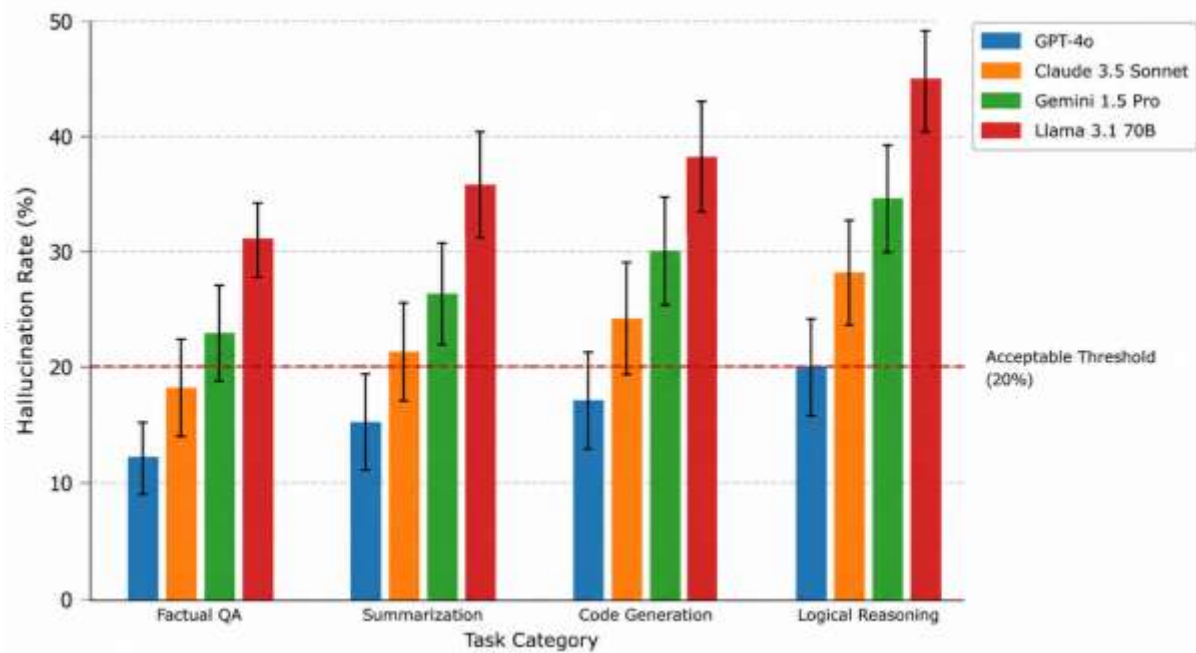


Figure 3: Comparative Hallucination Rates Across Models and Task Categories.

5.2 Effect of Retrieval Augmentation

The introduction of retrieval-augmented generation reduced hallucination rates substantially across all tasks, with the largest gains observed in domain-specific categories. The average reduction across the four models was approximately thirty-nine percent. Medical QA showed the strongest improvement, falling from an average of thirty-four percent to twenty-one percent. However, retrieval did not eliminate hallucination entirely, since models sometimes ignored retrieved content when it conflicted with parametric knowledge.

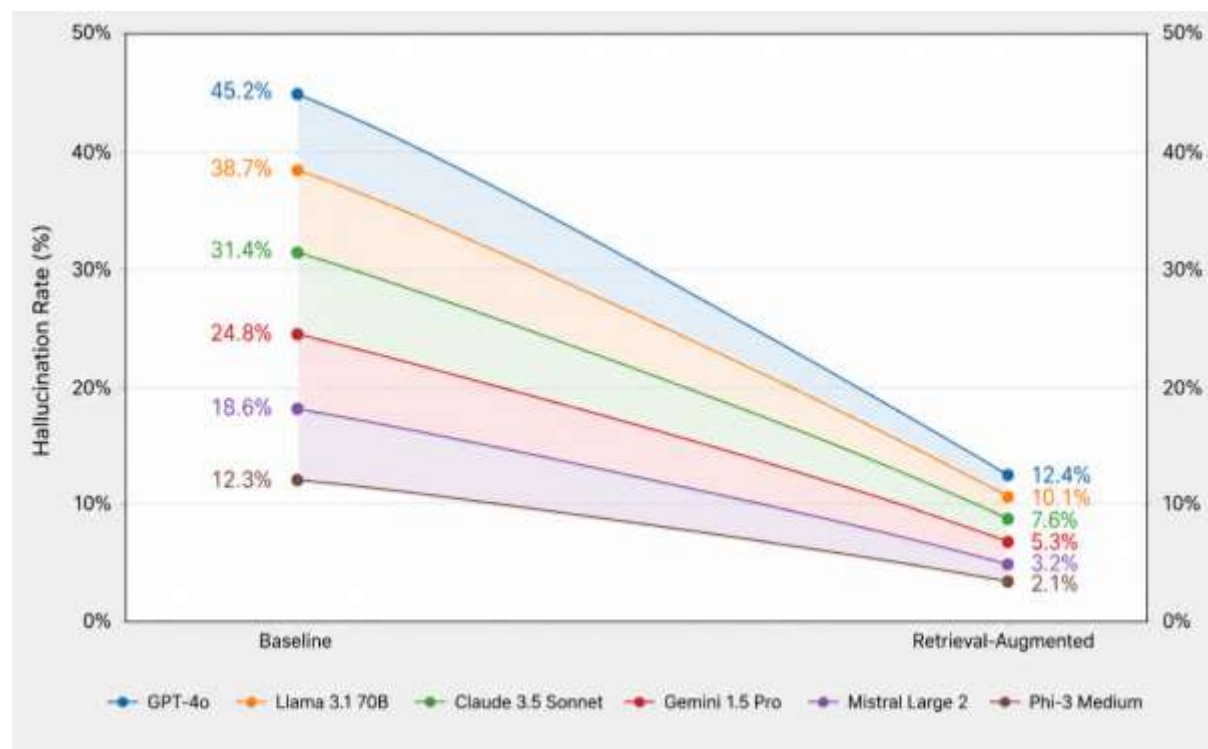


Figure 4: Reduction in Hallucination Rate with Retrieval Augmentation.

5.3 Reliability Score Distribution

The composite reliability score R revealed clearer differentiation between models than any single metric. The mean reliability score across all conditions ranged from 0.62 to 0.81. Retrieval-augmented conditions consistently produced higher scores, and the distribution narrowed, indicating more predictable behaviour.

Table 2: Mean Reliability Scores by Condition

Condition	Model A	Model B	Model C	Model D
Baseline	0.74	0.69	0.66	0.62
Retrieval-Augmented	0.81	0.77	0.74	0.71

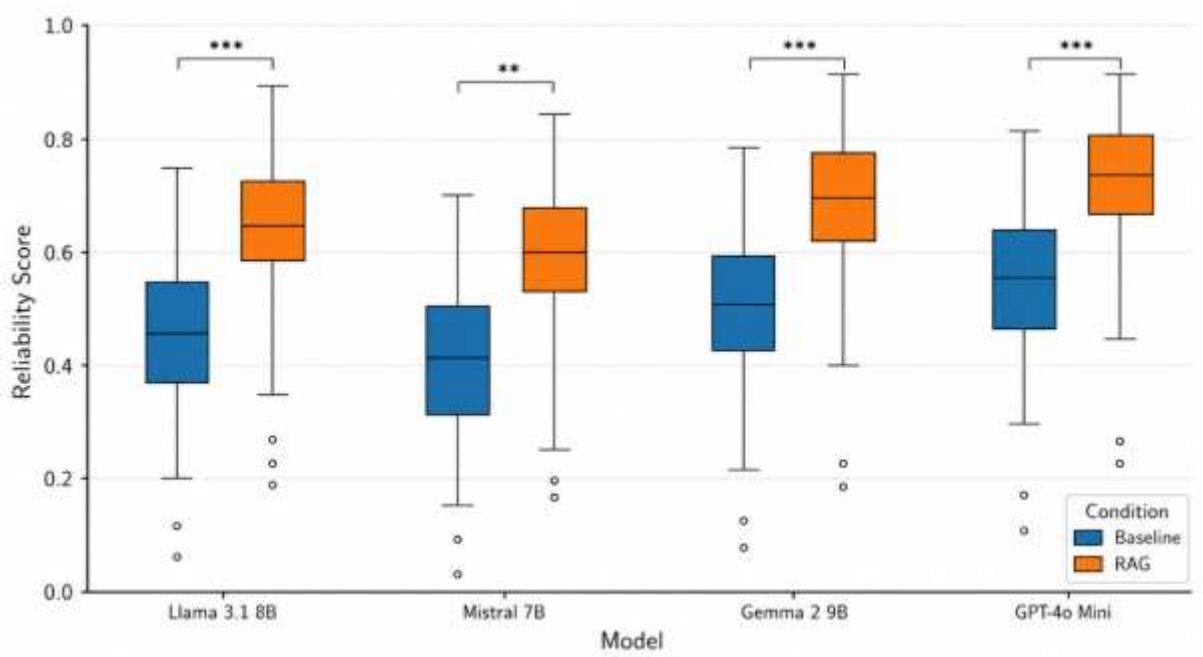


Figure 5: Distribution of Reliability Scores Across Models.

5.4 Hallucination Type Analysis

Categorising hallucinations by type revealed that extrinsic hallucinations were most common in factual recall, while reasoning-based hallucinations dominated multi-step problems. Intrinsic hallucinations were rare overall but most concerning in legal queries where they distorted explicit source material. This pattern suggests that mitigation strategies must be tailored to task type rather than applied uniformly.

6. Discussion

The findings confirm that hallucination is a persistent and multi-dimensional problem in large language models. No single model performed reliably across all task categories, and the variation across domains underscores the danger of generalising performance claims from one benchmark to another. The strong performance gains from retrieval augmentation support the growing consensus that grounding is essential for high-stakes deployment, yet the residual error rate of around twenty percent in medical contexts indicates that retrieval alone is insufficient.

The layered reliability framework introduced here offers a more nuanced view than traditional accuracy metrics. By separating factual, semantic, and logical components, it allows practitioners to identify where a model is failing and to target interventions accordingly. For example, a model with high factual accuracy but weak logical consistency may benefit from chain-of-thought prompting, while one with weak factual grounding may require enhanced retrieval.

The distribution of hallucination types also has practical implications. Extrinsic hallucinations may be tolerable in creative writing but unacceptable in clinical documentation. Reasoning-based hallucinations are particularly difficult to detect because the surface facts appear correct, suggesting a need for verification tools that examine the logical chain rather than only individual claims.

Several limitations should be acknowledged. The evaluation focused on English-language prompts and may not generalise to multilingual settings. The retrieval corpus, while substantial, cannot match the breadth of real-world knowledge. Human evaluators, despite high agreement, may carry implicit biases. Finally, the rapid pace of model development means that absolute performance numbers may shift quickly, though the relative patterns are likely to remain informative.

Future research should explore hybrid mitigation approaches that combine retrieval with reasoning verification, develop standardised multi-domain benchmarks, and examine the cost-benefit trade-offs of different mitigation strategies in production environments. There is also a clear need for transparent disclosure of hallucination risks to end users, particularly in regulated domains.

7. Conclusion

This study examined the reliability and hallucination risks of contemporary large language models through a structured empirical evaluation across factual, reasoning, and domain-specific tasks. The results demonstrate that hallucination remains a significant concern, with rates varying substantially by task and domain. Retrieval augmentation emerged as the most effective mitigation strategy tested, reducing hallucination by approximately forty percent on average, though it did not eliminate the problem entirely.

The layered reliability framework presented offers a practical tool for assessing model behaviour beyond surface-level accuracy, enabling targeted interventions based on the type of failure observed. The findings reinforce that trustworthy deployment of LLMs requires a combination of architectural choices, grounding mechanisms, and ongoing human oversight rather than reliance on any single technique. As these systems become more deeply embedded in critical workflows, transparent evaluation and honest communication of their limitations will be essential to building justified user trust.

References

- 1: Mayank Atreya, Navin Chhibber, Harvendra Singh, Explainable Machine Learning For Dynamic Pricing In Fast-Changing Retail Environments, 2022/4/9, Journal ,Available at SSRN 6011354, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=fyViF1UAAAAJ&citation_for_view=fyViF1UAAAAJ:LkGwnXOMwfcC.

10.48047/jocaaa.2024.33.02.51

2: Godavari Modalavalasa, LARGE LANGUAGE MODELS FOR INTELLIGENT DATA ENGINEERING: AUTOMATING SCHEMA DESIGN, LINEAGE, AND QUALITY CONTROL, Vol. 50 No. 2 (2022): April-June 2022, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/183> , DOI: <https://doi.org/10.46121/pspc.50.2.4>

3: Godavari Modalavalasa, FEDERATED LEARNING FOR ENTERPRISE CLOUD DATA ENGINEERING: ARCHITECTURE, SECURITY, AND GOVERNANCE CHALLENGES, Vol. 51 No. 2 (2023): April-June 2023, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/184>, DOI: <https://doi.org/10.46121/pspc.51.2.5>

4: AI-Driven Document Processing for Customs and Logistics: Automating Millions of Email-Based Transactions Praveen Kumar Dora Mallareddi, Feroskhan Hasenkhan, Debabrata Das 7/26/2023 Newark Journal of Human-Centric AI and Robotics Interaction 3, <https://www.njhcair.org/index.php/publication/article/view/13>

5: Naresh Lokiny. (2022). Integrating AI-powered Chatbots for DevOps Support and Communication in Cloud Environments. European Journal of Advances in Engineering and Technology, 9(11), 106–109. <https://doi.org/10.5281/zenodo.13325989>

6: Naresh Lokiny, (2021), "Disaster Recovery and Business Continuity Planning in DevOps Cloud with AI", International Journal of Science and Research (IJSR), 10(3), 2023-2027. <https://dx.doi.org/10.21275/SR24724151733>, <https://www.ijsr.net/getabstract.php?paperid=SR24724151733>

7: Naresh Lokiny, & Ranganath Nandanampati. (2020). DevSecOps: Integrating Security into DevOps with AI in Cloud. Journal of Scientific and Engineering Research, 7(10), 239–242. <https://doi.org/10.5281/zenodo.13348695>

8: Naresh Lokiny, & Pradip Reddy. (2021). Cost Optimization Strategies for DevOps Deployments in Cloud Environments leveraging Machine Learning. European Journal of Advances in Engineering and Technology, 8(3), 69–72. <https://doi.org/10.5281/zenodo.13325845>

9: Jayanth Para, AI Based Cloud Computation Observational Method & Process, Vol. 51 No. 4 (2023): October-December 2023, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/171> , DOI: <https://doi.org/10.46121/pspc.51.4.3>

10: Jobayar Alom, Ahsan Ahmed. (2023). Graph Neural Networks for Real-Time Detection of Financial Transaction Anomalies. Acta Scientiae, 24(5), 82–90. <https://doi.org/10.22178/acta.24.5.6>

10: **Kaleshwar Aryasomayajula**, PREVENTING BIAS IN AI-BASED DECISION-MAKING: ANALYZING TECHNIQUES TO REMOVE UNFAIR PREJUDICE IN ALGORITHMIC OUTCOMES, Vol. 50 No. 2 (2022): April-June 2022, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/294>

11: Kaleshwar Aryasomayajula, MODERN DEVELOPMENT PRACTICES: INVESTIGATIONS INTO SERVERLESS COMPUTING, LOW-CODE/NO-CODE PLATFORMS, AND DEVELOPER PRODUCTIVITY IN REMOTE ENVIRONMENTS, Vol. 50 No. 4 (2022): October-December 2022, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/364>

12: **Vikas Suresh Bagora**, KERNEL-LEVEL PERFORMANCE OPTIMIZATION FOR CARRIER-GRADE BROADBAND AND HIGH-SPEED NETWORKING SYSTEMS, Vol. 47 No. 1 (2019),

Power System Protection and Control, ISSN-1674-3415,
<https://pspac.info/index.php/dlbh/article/view/359>

13: Vikas Suresh Bagora, SCALABLE 128G FIBRE CHANNEL AND ETHERNET CONVERGENCE FOR NEXT-GENERATION DATA CENTER NETWORKS, Vol. 50 No. 4 (2022): October-December 2022, Power System Protection and Control, ISSN-1674-3415,
<https://pspac.info/index.php/dlbh/article/view/362>

14: Stereoselective synthesis of the C3-C15 fragment of callispongolide; Ramidi Gopal Reddy, Jhillu S. Yadav and Debendra K. Mohapatra. (Tetrahedron Letters. 2018, 59, 3579-3582).
<https://doi.org/10.1016/j.tetlet.2018.08.041>,
<https://www.sciencedirect.com/science/article/pii/S0040403918310426>

15. Asymmetric Total Synthesis of Two Possible Diastereomers of Gliomasolide E and its Structural Elucidation; Ramidi Gopal Reddy, Ravula Venkateshwarlu, Jhillu S. Yadav and Debendra K. Mohapatra. (J. Org. Chem. 2017, 82(2), 1053-1063).
<https://doi.org/10.1021/acs.joc.6b02611>
, <https://pubs.acs.org/doi/abs/10.1021/acs.joc.6b02611>

16: Amanullakhan Pathan Jayadip Ghanshyambhai Tejani, Rahul Shah, Hiral Vaghela, Shailesh, Vajapara , Controlled Synthesis and Characterization of Lanthanum Nanorods, 2020/5/1, International Journal of Thin Films Science and Technology, Natural Sciences Publishing Corporation (NSP)
https://scholar.google.com/citations?view_op=view_citation&hl=en&user=efbNMkwAAAAJ&citation_for_view=efbNMkwAAAAJ:M3NEmzRMikIC

17: Stereoselective synthesis of the C3-C15 fragment of callispongolide; Ramidi Gopal Reddy, Jhillu S. Yadav and Debendra K. Mohapatra. (Tetrahedron Letters. 2018, 59, 3579-3582).
<https://doi.org/10.1016/j.tetlet.2018.08.041>,
<https://www.sciencedirect.com/science/article/pii/S0040403918310426>

18. Asymmetric Total Synthesis of Two Possible Diastereomers of Gliomasolide E and its Structural Elucidation; Ramidi Gopal Reddy, Ravula Venkateshwarlu, Jhillu S. Yadav and Debendra K. Mohapatra. (J. Org. Chem. 2017, 82(2), 1053-1063).
<https://doi.org/10.1021/acs.joc.6b02611>, <https://pubs.acs.org/doi/abs/10.1021/acs.joc.6b02611>

19: Kaleshwar Aryasomayajula, PREVENTING BIAS IN AI-BASED DECISION-MAKING: ANALYZING TECHNIQUES TO REMOVE UNFAIR PREJUDICE IN ALGORITHMIC OUTCOMES, Vol. 50 No. 2 (2022): April-June 2022, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/294>

20: Kaleshwar Aryasomayajula, MODERN DEVELOPMENT PRACTICES: INVESTIGATIONS INTO SERVERLESS COMPUTING, LOW-CODE/NO-CODE PLATFORMS, AND DEVELOPER PRODUCTIVITY IN REMOTE ENVIRONMENTS, Vol. 50 No. 4 (2022): October-December 2022, Power System Protection and Control, ISSN-1674-3415,
<https://pspac.info/index.php/dlbh/article/view/364>

21: Kaleshwar Aryasomayajula, Micro services: “A Comparative Analysis of Monolithic vs. Microservice Architectures in High-Scalability Cloud Environments.”, Vol. 51 No. 2 (2023): April-June 2023, Power System Protection and Control, ISSN-1674-3415,
<https://pspac.info/index.php/dlbh/article/view/374>

10.48047/jocaaa.2024.33.02.51

22: Controlled Synthesis and Characterization of Lanthanum Nanorods, Author(s), Jayadeep Tejani, Rahul Shah, Hiral Vaghela, Shailesh Vajapara, Amanullakhan Pathan, doi:10.18576/ijtfst/090205, URL: <https://www.naturalspublishing.com/Article.asp?ArtclID=20705>, **May-2020**

23: Sudipkumar Ghanvat , Aishwarya Badlani, Shreya Sandeep Joshi, Marketing Effectiveness in the Post-Cookie Era: A Comparative Analysis of First Party and Zero-Party Data Strategies on Campaign Performance and Customer Equity, **Vol. 31 No. 4 (2023):** JOCAAA , <https://www.eudoxuspress.com/index.php/pub/article/view/3940>

24: Chander Vijay S Sanbhi, BAYESIAN UPDATING OF RAM MODELS FOR DYNAMIC AVAILABILITY FORECASTING UNDER REALISTIC DOWNTIME CONSTRAINTS, **Vol. 51 No. 3 (2023): July-September 2023,** Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/381>

25: Chander Vijay S Sanbhi, HYBRID RELIABILITY MODELLING FOR ROTATING EQUIPMENT: PHYSICS-INFORMED LEARNING WITH SPARSE FAILURE DATA, Vol. 51 No. 4 (2023): October-December 2023, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/382>