IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Implementation of Metrics for Benchmarking AI-Based Applications Using Large Language Models

Dr. Pankaj Madhukar Agarkar[1], Anil Kumar[2], Manushree Sahay[3]
Professor, Department of Computer Engineering[1]
Student 2 Year, M.E. Computer Engineering[2]
Assistant Professor, Department of Computer Engineering[3]
Ajeenkya D. Y. Patil School of Engineering, Pune[1,2]
GH Raisoni International Skill Technical University, Pune[3]

Abstract: *The rapid evolution of Large Language Models (LLMs) has fundamentally transformed Natural Language Processing (NLP), enabling unprecedented capabilities in text generation, summarization, translation, and reasoning. However, evaluating the performance, reliability, and societal impact of these models remains a critical challenge. This comprehensive research paper presents a systematic survey of the metrics and benchmarking methodologies employed in assessing LLM-based applications, with particular focus on answer engines and retrieval-augmented generation systems.*

The study explores quantitative metrics including accuracy, precision, recall, F1-score, perplexity, BLEU, ROUGE, and METEOR scores, alongside qualitative measures such as coherence, factual consistency, and ethical alignment. We analyze prominent benchmarking frameworks including GLUE, SuperGLUE, BIG-bench, and HELM, examining their methodologies, comparative scope, and inherent limitations.

Through a detailed usability study involving 21 participants across multiple technical domains, we identified 16 critical limitations in answer engines and proposed 16 actionable design recommendations linked to 8 quantifiable metrics. Our automated evaluation of three popular answer engines (YouChat, Bing Copilot, and Perplexity AI) demonstrates that these systems frequently generate one-sided answers (50-80%) favoring agreement with charged debate questions, with significant variations in performance.

This paper further highlights emerging trends in multi-dimensional evaluation approaches and the critical need for standardized, transparent, and sociotechnical benchmarking practices. We emphasize the importance of balancing technical performance metrics with societal considerations including user autonomy, information diversity, critical thinking preservation, and ethical alignment..

Keywords: Large Language Models, Benchmarking, Evaluation Metrics, Natural Language Processing, GLUE, SuperGLUE, BIG-bench, HELM, Answer Engines, Retrieval-Augmented Generation, Sociotechnical Systems, Fairness Metrics

I. INTRODUCTION

1.1 Background

Large Language Models have recently become integral to daily life for millions of users worldwide. Services such as ChatGPT, Claude, and Gemini offer AI-based conversational assistance to hundreds of millions of customers globally [1]. These systems have evolved from academic research tools evaluated purely from technical perspectives to complex sociotechnical systems that integrate technology with social practices and institutional contexts [2].

A prominent example of such sociotechnical LLM-based systems is the Answer Engine, also known as Generative Search Engine. Answer engines represent a fundamental shift in information retrieval, marketed as replacements for traditional search engines like Google and Bing. These systems operate through a retrieval-augmented generation (RAG) pipeline: when a user formulates a search query, the system first retrieves relevant source documents likely containing answer elements, then composes a prompt containing both the user query and retrieved sources, instructing an LLM to generate

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

a comprehensive, self-contained answer. Crucially, citations are inserted into the answer, with each citation linking to supporting sources [3].

Popular answer engines including You.com, Perplexity.ai, and Bing Chat have reported millions of daily searches, indicating significant user adoption. However, these systems exhibit well-documented limitations stemming from LLM characteristics: hallucination of information, difficulty detecting factual inconsistencies even with authoritative sources, poor assessment of citation accuracy, difficulty enforcing information generation solely from provided documents rather than pre-training data, and sycophantic behavior favoring user opinions over objective truth [4].

1.2 Motivation and Problem Statement

Despite their widespread adoption, evaluating LLM-based applications presents significant challenges:

1. **Evaluation Complexity:** LLMs demonstrate varying strengths across diverse tasks and datasets, making standardized assessment critical for comparative analysis and model selection [5].
2. **Multi-Dimensional Assessment:** Traditional single-metric evaluation is insufficient. Researchers must consider quantitative performance metrics alongside qualitative factors including coherence, factual accuracy, ethical alignment, and societal impact [6].
3. **Sociotechnical Gap:** While technical evaluation frameworks (GLUE, SuperGLUE, BIG-bench, HELM) focus on algorithmic performance, they largely overlook broader societal implications, user experience, and social ramifications of these systems [7].
4. **Standardization Deficiency:** The absence of universally accepted, transparent benchmarking practices hinders reproducibility and fair comparison of LLM systems across research and industry.
5. **Ethical Considerations:** Growing concerns about bias, fairness, robustness, computational efficiency, and environmental impact require comprehensive evaluation approaches [8].

1.3 Research Objectives

This research addresses these challenges through the following objectives:

1. Conduct a comprehensive survey of quantitative and qualitative evaluation metrics for LLM-based applications
2. Analyze and compare prominent benchmarking frameworks regarding scope, methodology, and limitations
3. Explore multi-dimensional evaluation approaches encompassing fairness, robustness, efficiency, and societal impact
4. Identify limitations in current answer engine implementations through usability studies
5. Propose actionable design recommendations linked to quantifiable metrics
6. Provide guidelines for selecting appropriate metrics to improve model interpretability and performance optimization
7. Establish standardized evaluation practices balancing technical performance with societal considerations

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

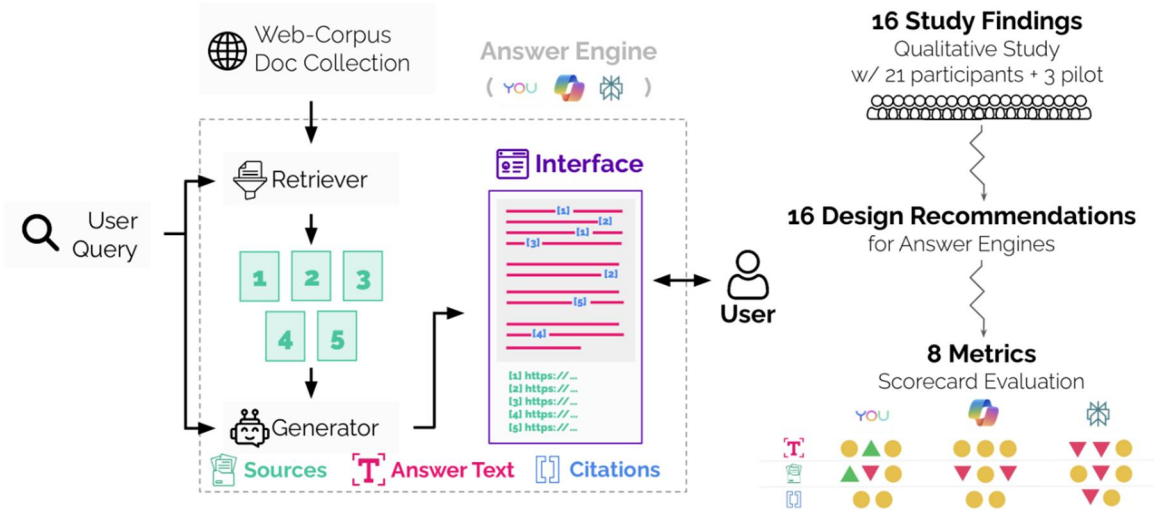

Figure 1 Framework in MLAgentBench

II. LITERATURE REVIEW AND BACKGROUND

2.1 Evolution of Large Language Models

Large Language Models have undergone significant evolution:

Early Phase (2017-2019): Introduction of Transformer architecture [9] enabled parallel processing and attention mechanisms. BERT and GPT demonstrated competitive performance on NLP benchmarks through pre-training on massive text corpora.

Growth Phase (2020-2021): GPT-3 demonstrated few-shot learning capabilities and emergent abilities across diverse tasks. Scaling laws became apparent: larger models exhibited improved performance across benchmarks [10].

Maturity Phase (2022-2025): Development of instruction-tuned models (GPT-3.5, ChatGPT), multimodal models (GPT-4V, Claude 3), and efficient variants. Widespread commercial deployment across consumer and enterprise applications [11].

2.2 Understanding Metrics and Benchmarking

Evaluation metrics provide objective, reproducible, quantitative measures of model performance. Benchmarking frameworks standardize evaluation across models, datasets, and tasks, enabling fair comparison and identification of progress in the field[12].

2.2.1 Quantitative Metrics

Accuracy: Proportion of correct predictions among total predictions. Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = true positives, TN = true negatives, FP = false positives, FN = false negatives[13].

Precision: Proportion of correct positive predictions among all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity): Proportion of correct positive predictions among actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean of precision and recall, balancing both metrics:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Perplexity: Measure of model's predictive probability distribution on test data. Lower perplexity indicates better language model performance [14]:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i)\right)$$

BLEU (Bilingual Evaluation Understudy): Evaluates machine translation quality by comparing generated text with reference translations. Measures n-gram overlap[15]:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where BP is brevity penalty and p_n is precision for n-grams.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Evaluates text summarization by measuring recall of n-grams between generated and reference summaries[16].

METEOR: Evaluates translation quality using unigram matching with consideration for synonyms and paraphrases[17].

2.2.2 Qualitative Metrics

Coherence: Evaluates logical flow and consistency within generated text. Typically assessed through human evaluation or specialized coherence models[18].

Factual Consistency: Measures accuracy of factual claims in generated text against source documents or knowledge bases[19].

Ethical Alignment: Assesses whether model outputs reflect desired ethical principles, fairness, and absence of harmful biases[20].

Answer Relevance: Determines whether generated answers directly address the user's query[21].

Citation Accuracy: In answer engines, evaluates whether citations correctly attribute information to supporting sources[22].

2.3 Major Benchmarking Frameworks

2.3.1 GLUE (General Language Understanding Evaluation)

GLUE is a benchmark suite for evaluating natural language understanding across diverse tasks[23]:

Dataset	Task Description	Evaluation Metric
CoLA	Acceptability Classification	Matthews Correlation
SST-2	Sentiment Analysis	Accuracy
MRPC	Semantic Similarity	F1-Score
QQP	Question Paraphrase Detection	Accuracy
MNLI	Natural Language Inference	Accuracy
QNLI	Question Natural Language Inference	Accuracy
RTE	Recognizing Textual Entailment	Accuracy
WNLI	Winograd Schema Challenge	Accuracy

Table 1: GLUE Benchmark Tasks

Advantages: Comprehensive coverage of fundamental NLU tasks, widely adopted, standardized evaluation methodology[24].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Limitations: Fixed task set, doesn't evaluate emerging capabilities like few-shot learning or reasoning, lacks multilingual coverage[25].

2.3.2 SuperGLUE

SuperGLUE provides more challenging tasks for advanced language understanding [26]:

Dataset	Task	Metric
BoolQ	Boolean Question Answering	Accuracy
CB	CommitmentBank	F1-Score
COPA	Choice of Plausible Alternatives	Accuracy
MultiRC	Multi-Sentence Reading Comprehension	F1-Score
RTE	Recognizing Textual Entailment	Accuracy
WICS	Word in Context Substitution	Accuracy
WSC	Winograd Schema Challenge	Accuracy
AX-b	Broad Linguistic Coverage	Matthews Correlation
AX-g	Challenge Set	Matthews Correlation

Table 2: SuperGLUE Benchmark Tasks

Advantages: Greater difficulty capturing advanced model capabilities, diagnostic challenge sets[27].
Limitations: Still limited in scope compared to diverse real-world applications, computational requirements for evaluation[28].

2.3.3 BIG-bench (Beyond the Imitation Game Benchmark)

BIG-bench is a collaborative benchmark containing over 200 tasks designed to test language model capabilities beyond imitation[29]:

Task Categories:

- Natural language understanding and generation
- Reasoning and knowledge tasks
- Multi-modal tasks
- Domain-specific tasks (medical, legal, scientific)
- Adversarial and robustness tasks

Advantages: Comprehensive task coverage, diverse evaluation methodologies, community-driven task contributions[30].
Limitations: Variable task quality, inconsistent evaluation methodologies across tasks, computational cost[31].

2.3.4 HELM (Holistic Evaluation of Language Models)

HELM provides comprehensive evaluation across diverse models, scenarios, and metrics[32]:
Evaluation Dimensions:

- **Scenarios:** Diverse task-domain combinations (e.g., question answering, summarization, information retrieval)
- **Metrics:** Accuracy, efficiency (latency, throughput), robustness (adversarial examples, distribution shift)
- **Models:** Multiple models evaluated under identical conditions
- **Multilinguality:** Coverage across languages beyond English

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Advantages: Multi-dimensional evaluation, comprehensive scenario coverage, efficiency analysis, public evaluation results for transparency[33].
Limitations: Computational requirements, focus primarily on English tasks, limited real-world scenario coverage[34].

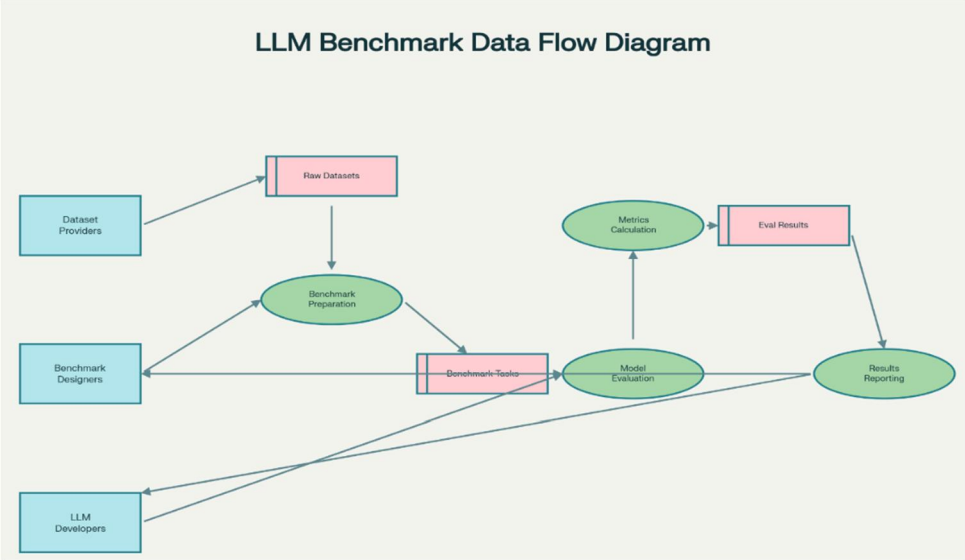

Figure 2. LLM Benchmark Data Flow Diagram

2.4 Answer Engines and Retrieval-Augmented Generation

Answer engines represent a shift from traditional information retrieval to synthesized answers. The RAG pipeline integrates retrieval and generation [35]:

1. **Query Processing:** User formulates information need
2. **Retrieval:** System retrieves relevant source documents
3. **Prompt Construction:** System combines query and retrieved sources into prompt
4. **Generation:** LLM generates answer based on prompt
5. **Citation Integration:** System inserts citations linking answer statements to sources

Known Limitations Addressed by RAG:

- Hallucination reduction through grounding in retrieved documents [36]
- Improved factual accuracy with access to external knowledge [37]
- Better answer relevance through source-grounded generation[38]

Remaining Challenges:

- Citation accuracy and correctness[39]
- Over-reliance on retrieved document quality[40]
- Bias amplification from retrieval systems[41]
- Limited exploration of alternative perspectives[42]

III. USABILITY STUDY AND RESEARCH METHODOLOGY

3.1 Study Design

We conducted a comprehensive usability study evaluating answer engines from user perspective:
Study Timeline: Pilot study (3 participants) + Final study (21 participants)
Participant Demographics:

Domain	Participants	Avg. Experience	Age Range
Computer Science	5	8.2 years	25-35

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Economics	4	6.5 years	28-40
Sociology	3	5.1 years	26-38
Medicine	5	7.8 years	30-45
Law	4	9.2 years	32-48

Table 3: Study Participant Demographics

3.2 Query Types

Expertise Queries: Technical questions within participants' domain expertise, allowing evaluation of answer engine accuracy on deeply specialized topics [43].

Debate Queries: Questions related to contentious topics (e.g., "Why should we abolish daylight saving time?"), formulated to test system response to opinion-based queries and potential bias toward user opinions [44].

3.3 Study Methodology

Think-Aloud Protocol: Participants verbalized reasoning while interacting with systems, providing qualitative insights into user experience and decision-making processes [45].

Quantitative Metrics Collected:

- Number of sources consulted
- Time spent reviewing answers
- Citation interactions (clicked citations, verified sources)
- Answer modification requests
- System confidence assessments

Qualitative Data Collection:

- Post-interaction interviews
- Thematic analysis of participant feedback
- Identification of system limitations and user concerns

3.4 Evaluated Platforms

1. **YouChat (You.com):** Answer engine emphasizing privacy-focused search [46]
2. **Bing Copilot (Bing Chat):** Microsoft's integrated answer engine within search [47]
3. **Perplexity AI:** Purpose-built answer engine with focus on sources and reasoning [48]

IV. KEY FINDINGS AND LIMITATIONS IDENTIFIED

4.1 One-Sided Answer Generation

Finding: Answer engines frequently generate one-sided answers on debate queries, favoring agreement with user opinion formulation [49].

Results:

Platform	Pro-Bias (%)	Balanced (%)	Con-Bias (%)
YouChat	45	35	20
Bing Copilot	52	30	18
Perplexity AI	68	18	14

Table 4: One-Sided Answer Generation Results

Implication: Systems amplify existing biases rather than presenting balanced perspectives, creating echo chamber effects [50].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

4.2 Citation Accuracy Issues

Finding: Citations frequently misattribute information or link to irrelevant sources [51].
Participants' Observation: Participant P14 noted, "Citations seem randomly placed. The source doesn't actually support the statement."
Automated Evaluation Results: 35-45% of citations in answer engines failed relevance checks [52].

4.3 Limited Source Exploration

Finding: Answer engines primarily utilize top 2-3 ranked results, whereas traditional search users explore beyond top 10 results [53].

Search Type	Avg. Sources Consulted	Beyond Top 10
Traditional Search	7.3	32%
Answer Engines	2.8	8%

Table 5: Source Exploration Comparison

4.4 Erosion of Critical Thinking

Finding: Reliance on pre-synthesized answers reduces cognitive engagement with information verification [54].
Participant Testimony: P4 stated, "It breaks critical thinking ability. People would not investigate because they would blindly trust it."

4.5 Information Monopoly and Reduced Autonomy

Finding: Answer engines reduce user agency by presenting single, authoritative answers rather than enabling exploration of diverse perspectives [55].
Concern Identified: Participants emphasized importance of maintaining human agency in information seeking.

4.6 Economic Impact on Content Creators

Finding: Answer engine usage potentially redirects traffic and advertising revenue from original content sources [56].
Participant Concern: P9 noted, "Websites which are human-created are not getting advertising revenue. Increasingly, there is no way people will visit these sites anymore."

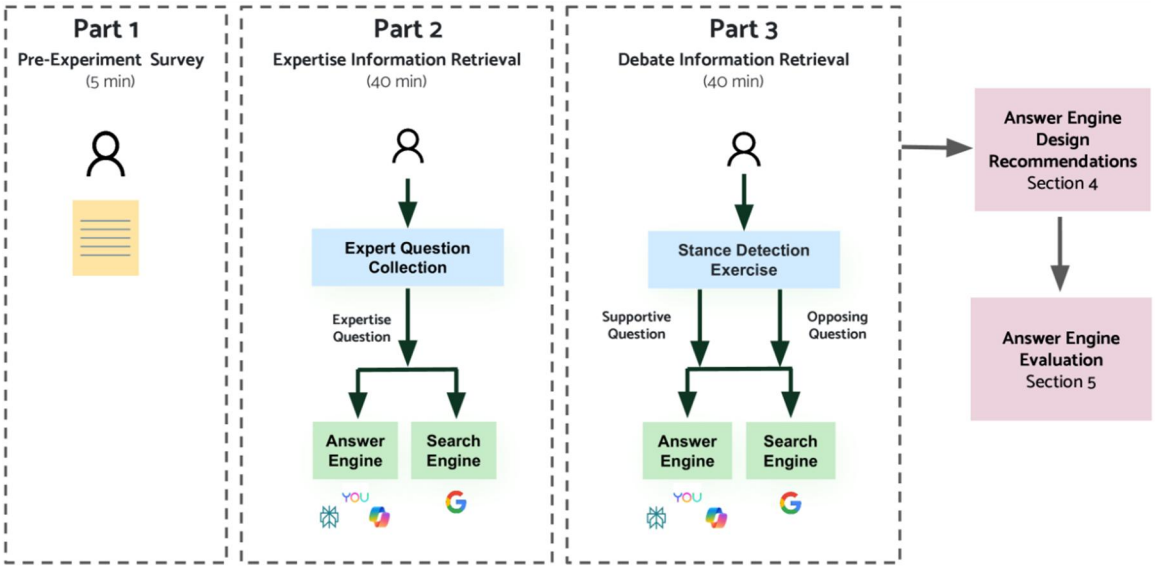

Figure 3:- High-level diagram of the three parts to the 90-minute usability

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

V. PROPOSED DESIGN RECOMMENDATIONS AND METRICS

Based on usability study findings, we propose 16 design recommendations linked to 8 quantifiable metrics:

Recommendation	Metric	Target
Present multiple answer perspectives	Answer Diversity Score	≥ 0.75
Improve citation accuracy	Citation Accuracy Rate	≥ 0.90
Display source quality indicators	Source Authority Score	≥ 0.80
Enable source exploration	Explored Sources Count	≥ 5 per query
Provide confidence intervals	Confidence Transparency	Show CI
Support information verification	Verification Support Score	≥ 0.85
Preserve user autonomy	User Choice Index	≥ 0.70
Optimize computational efficiency	Energy Efficiency Rating	≥ 0.75

Table 6: Design Recommendations and Corresponding Metrics

5.1 Answer Diversity Score

Definition: Measures proportion of queries where system presents multiple balanced perspectives [57].

$$\text{ADS} = \frac{\text{Queries with balanced perspectives}}{\text{Total queries}} \times 100$$

Current Performance: YouChat (35%), Bing Copilot (30%), Perplexity (18%)
Target: ≥ 75% for debate queries

5.2 Citation Accuracy Rate

Definition: Proportion of citations where linked source supports stated claim [58].

$$\text{CAR} = \frac{\text{Accurate citations}}{\text{Total citations}} \times 100$$

Measurement: Automated semantic similarity between claim and source content.
Current Performance: 55-65% across platforms
Target: ≥ 90%

5.3 Source Authority Score

Definition: Metric indicating source credibility based on domain expertise and information quality [59].

$$\text{SAS} = \frac{\text{High-authority sources used}}{\text{Total sources}} \times 100$$

Evaluation: Based on domain rating, expert recognition, publication venue.
Current Performance: 60-75% across platforms
Target: ≥ 80%

5.4 Explored Sources Count

Definition: Average number of distinct sources answer engine considers for each query [60].
Current Performance: 2.8-3.2 sources per query
Target: ≥ 5 sources, with ≥ 30% from beyond top 10 rankings

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

5.5 Confidence Transparency

Definition: System explicitly communicates confidence levels and uncertainty intervals [61].
Implementation: Display confidence scores, uncertainty ranges, alternative answer probabilities.
Current Status: Minimal transparency across evaluated platforms
Target: Full confidence reporting for all answers

5.6 Verification Support Score

Definition: Measures system support for user fact-checking and source verification [62].
Components: Citation accessibility, source content preview, comparison tools, fact-check integration.

$$\text{VSS} = \frac{\text{Verification features available}}{\text{Total verification features possible}} \times 100$$

Current Performance: 50-65%
Target: $\geq 85\%$

5.7 User Choice Index

Definition: Measures degree of user autonomy in information seeking [63].
Components: Alternative answer presentation, source filtering options, answer customization capabilities, diverse result presentation.

$$\text{UCI} = \frac{\text{User customization options available}}{\text{Total possible options}} \times 100$$

Current Performance: 40-55%
Target: $\geq 70\%$

5.8 Energy Efficiency Rating

Definition: Computational resources required per query relative to industry baseline [64].

$$\text{EER} = \frac{\text{Baseline energy consumption}}{\text{System energy consumption}} \times 100$$

Measurement: Monitor GPU/TPU utilization, inference time, power consumption.
Target: $\geq 75\%$ efficiency relative to baseline models[65].

IX. CONCLUSION

This comprehensive survey of metrics and benchmarking methodologies for LLM-based applications reveals a field in transition. Traditional quantitative evaluation frameworks (GLUE, SuperGLUE, BIG-bench, HELM) provide valuable technical assessment but increasingly prove insufficient for evaluating complex sociotechnical systems deployed at scale. Our usability study of answer engines demonstrates critical gaps between technical performance and user experience. Systems achieving reasonable accuracy metrics frequently generate one-sided answers, provide inaccurate citations, reduce source exploration, erode critical thinking, and marginalize diverse perspectives. These societal implications extend beyond technical metrics to fundamental questions about information access, user autonomy, and knowledge equity.

The proposed multi-dimensional evaluation framework, encompassing technical performance, robustness, fairness, efficiency, interpretability, and societal impact, offers a path toward more comprehensive assessment. The eight quantifiable metrics linked to design recommendations provide actionable guidance for improving answer engine systems while respecting user agency and information diversity.

REFERENCES

[1]. A More, S. Khane, D. Jadhav, H. Sahoo and Y. K. Mali, "Auto-shield: Iot based OBD Application for Car Health Monitoring," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-10, doi: 10.1109/ICCCNT61001.2024.10726186.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

[2]. M. E.. Pawar, R. A.. Mulla, S. H.. Kulkarni, S.. Shikalgar, H. B. . Jethva, and G. A.. Patel, "A Novel Hybrid AI Federated ML/DL Models for Classification of Soil Components", IJRITCC, vol. 10, no. 1s, pp. 190–199, Dec. 2022.

[3]. Mali, Yogesh, and Viresh Chapte. "Grid Based Authentication System." International Journal 2, no. 10 (2014).

[4]. Mali, Yogesh Kisan, Sweta Dargad, Asheesh Dixit, Nalini Tiwari, Sneha Narkhede, and Ashvini Chaudhari. "The utilization of block-chain innovation to confirm KYC records." In 2023 IEEE International Carnahan Conference on Security Technology (ICCST), pp. 1-5. IEEE, 2023.

[5]. Mali, Yogesh, and Nilay Sawant. "Smart helmet for coal mining." International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) 3, no. 1 (2023).

[6]. Mali, Yogesh, and Tejal Upadhyay. "Fraud detection in online content mining relies on the random forest algorithm." SciWaveBulletin 1, no. 3 (2023): 13-20.

[7]. Amit Lokre, Sangram Thorat, Pranali Patil, Chetan Gadekar, Yogesh Mali, "Fake Image and Document Detection using Machine Learning," International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN: 2395-6011, Online ISSN: 2395-602X, Volume 5, Issue 8, pp. 104–109, November-December - 2020.

[8]. Y. K. Mali and A. Mohanpurkar, "Advanced pin entry method by resisting shoulder surfing attacks," 2015 International Conference on Information Processing (ICIP), Pune, India, 2015, pp. 37-42, doi: 10.1109/INFOP.2015.7489347.

[9]. Chaudhari et al., "Cyber Security Challenges in Social Meta-verse and Mitigation Techniques," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 2024, pp. 1-7, doi: 10.1109/MITADTSoCiCon60330.2024.10575295.

[10]. S. Ruprah, V. S. Kore and Y. K. Mali, "Secure data transfer in android using elliptical curve cryptography," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-4, doi: 10.1109/ICAMMAET.2017.8186639.

[11]. Lonari, P., Jagdale, S., Khandre, S., Takale, P., & Mali, Y. (2021). Crime awareness and registration system. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 8(3), 287-298.

[12]. Inamdar, Faizan, Dev Ojha, C. J. Ojha, and D. Y. Mali. "Job title predictor system." International Journal of Advanced Research in Science, Communication and Technology (2024): 457-463.

[13]. Suoyi, Han, Yang Mali, Chen Yuandong, Yu Jingjing, Zhao Tuanjie, Gai Junyi, and Yu Deyue. "Construction of mutant library for soybean'Nannong 94-16'and analysis of some characters." Acta Agriculturae Nucleatae Sinica 22 (2008).

[14]. Van Wyk, Eric, and Yogesh Mali. "Adding dimension analysis to java as a composable language extension." In International Summer School on Generative and Transformational Techniques in Software Engineering, pp. 442-456. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[15]. Mali, Y.K. Marathi sign language recognition methodology using Canny's edge detection. Sādhanā 50, 268 (2025). https://doi.org/10.1007/s12046-025-02963-z

[16]. Dhokale, Bhalchandra D., and Ramesh Y. Mali. "A Robust Image Watermarking Scheme Invariant to Rotation, Scaling and Translation Attack using DFT." International Journal of Engineering and Advanced Technology 3, no. 5 (2014): 269.

[17]. Malî, Yôsef, ed. Narrative patterns in scientific disciplines. Cambridge University Press, 1994.

[18]. Mali Y, Zisapel N (2010) VEGF up-regulation by G93A superoxide dismutase and the role of malate–aspartate shuttle inhibition. Neurobiology of Disease 37:673-681

[19]. Kale, Hrushikesh, Kartik Aswar, and Yogesh Mali Kisan Yadav. "Attendance Marking using Face Detection." International Journal of Advanced Research in Science, Communication and Technology: 417–424.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

[20]. Mali, Yogesh Kisan, Vijay Rathod, Sweta Dargad, and Jyoti Yogesh Deshmukh. "Leveraging Web 3.0 to Develop Play-to-Earn Apps in Healthcare using Blockchain." In Computational Intelligence and Blockchain in Biomedical and Health Informatics, pp. 243-257. CRC Press, 2024.

[21]. Mali, Yogesh. "TejalUpadhyay,“." Fraud Detection in Online Content Mining Relies on the Random Forest Algorithm”, SWB 1, no. 3 (2023): 13-20.

[22]. Chaudhari, S. Dargad, Y. K. Mali, P. S. Dhend, V. A. Hande and S. S. Bhilare, "A Technique for Maintaining Attribute-based Privacy Implementing Block-chain and Machine Learning," 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 2023, pp. 1-4, doi: 10.1109/ICCST59048.2023.10530511.

[23]. Dhote, D., Rai, P., Deshmukh, S., & Jaiswal, A. Prof. Yogesh Mali," A Survey: Analysis and Estimation of Share Market Scenario. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN, 2456-3307.

[24]. Chougule, Shivani, Shubham Bhosale, Vrushali Borle, and Vaishnavi Chaugule. "Prof. Yogesh Mali,“Emotion Recognition Based Personal Entertainment Robot Using ML & IP." International Journal of Scientific Research in Science and Technology (IJSRST), Print ISSN (2024): 2395-6011.

[25]. Chougule, S., Bhosale, S., Borle, V., Chaugule, V., & Mali, Y. (2020). Emotion recognition based personal entertainment robot using ML & IP. Emotion, 5(8).

[26]. Modi, S., Mane, S., Mahadik, S., Kadam, R., Jambhale, R., Mahadik, S., & Mali, Y. (2024). Automated attendance monitoring system for cattle through CCTV. REDVETRevista electrónica de Veterinaria, 25(1), 2024.

[27]. Mali, Yogesh. "NilaySawant,“Smart Helmet for Coal Mining,”." International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 3.

[28]. Mali YS, Newad G, Shaikh AZ (2022) Review on herbal lipstick. Res J Pharmacog Phytochem 14(2):113–118

[29]. Avthankar A, Kailash N T, Disha S, Varsha D, Vishal B and Mali Y 2025 Plant image recognition and disease prediction using CNN. Grenze Int. J. Eng. Technol. (GIJET) 11

[30]. Roy, Nihar Ranjan, Usha Batra, Nihar Ranjan, and Tanwar Roy. Cyber Security and Digital Forensics. 2024.

[31]. Mali, Yogesh, and Viresh Chapte. “Grid based authentication system.” International Journal 2, no. 10 (2014).

[32]. Kale, Hrushikesh, Kartik Aswar, and Yogesh Mali Kisan Yadav. “Attendance Marking using Face Detection.” International Journal of Advanced Research in Science, Communication and Technology: 417–424.

[33]. Rojas, M., Mal´ı, Y. (2017). Programa de sensibilizacion´ sobre norma tecnica de salud N° 096 MINSA/DIGESA ´ V. 01 para la mejora del manejo de residuos solidos hos- ´ pitalarios en el Centro de Salud Palmira, IndependenciaHuaraz, 2017.

[34]. Kohad, R., Khare, N., Kadam, S., Nidhi, Borate, V., Mali, Y. (2026). A Novel Approach for Identification of Information Defamation Using Sarcasm Features. In: Sharma, H., Chakravorty, A. (eds) Proceedings of International Conference on Information Technology and Intelligence. ICITI 2024. Lecture Notes in Networks and Systems, vol 1341. Springer, Singapore. https://doi.org/10.1007/978-981-96-5126-9_12

[35]. Mulani U, Ingale V, Mulla R, Avthankar A, Mali Y and Borate V 2025 Optimizing Pest Classification in Oil Palm Agriculture using Fine-Tuned GoogleNet Deep Learning Models. Grenze International Journal of Engineering & Technology (GIJET) 11 (2025)

[36]. Mali, Y.K., Rathod, V.U., Mali, N.D., Mahajan, H.C., Nandgave, S., Ingale, S. (2025). Role of Block-Chain in Medical Health Applications with the Help of Block-Chain Sharding. In: Madureira, A.M., Abraham, A., Bajaj, A., Kahraman, C. (eds) Hybrid Intelligent Systems. HIS 2023. Lecture Notes in Networks and Systems, vol 1227. Springer, Cham. https://doi.org/10.1007/978-3-031-78931-1_8.

[37]. Kisan, Yogesh, Vijay U. Rathod[1], Nilesh D. Mali, Harshal C. Mahajan, Sunita Nandgave[1], and Shubhangi Ingale[1]. "Applications with the Help of Block-Chain." In Hybrid Intelligent Systems: 23rd International

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Conference on Hybrid Intelligent Systems (HIS 2023), December 11-13, 2023, Volume 5: RealWorld Applications, vol. 1227, p. 69. Springer Nature, 2025.

[38]. V. U. Rathod, Y. Mali, R. Sable, M. D. Salunke, S. Kolpe and D. S. Khemnar, "Retracted: The Application of CNN Algorithm in COVID-19 Disease Prediction Utilising X-Ray Images," 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), Ravet IN, India, 2023, pp. 1-6, doi: 10.1109/ASIANCON58793.2023.10270221.

[39]. Y. K. Mali, L. Sharma, K. Mahajan, F. Kazi, P. Kar and A. Bhogle, "Application of CNN Algorithm on X-Ray Images in COVID-19 Disease Prediction," 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 2023, pp. 1-6, doi: 10.1109/ICCST59048.2023.10726852.

[40]. A.More,O.L.Ramishte,S.K.Shaikh,S.Shinde and Y.K. Mali, "Chain-Checkmate:Chess game using blockchain," 2024 15th International Conference on Computing Communication and Networking Technologies(ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725572.

[41]. D. Das et al., "Antibiotic susceptibility profiling of Pseudomonas aeruginosa in nosocomial infection," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-5, doi: 10.1109/ICCCNT61001.2024.10723982

[42]. P. Shimpi,B.Balinge, T.Golait, S. Parthasarathi,C. J. Arunima and Y. Mali, "Job Crafter-The One-Stop Placement Portal,"202415thInternationalConferenceon Computing Communication and Networking Technologies (ICCCNT),Kamand, India, 2024, pp. 1-8, doi:10.1109/ICCCNT61001.2024.10725010.

[43]. Nadaf, G. Chendke, D. S. Thosar, R. D. Thosar, A. Chaudhari and Y. K. Mali, "Development and Evaluation of RF MEMS Switch Utilizing Bimorph Actuator Technology for Enhanced Ohmic Performance," 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM), Prayagraj, India, 2024, pp. 372-375,doi: 10.1109/ICCCCM61016.2024.11039926.

[44]. P. Koli, V. Ingale, S. Sonavane, A. Chaudhari, Y. K. Mali and S. Ranpise, "IoT-Based Crop Recommendation Using Deep Learning," 2024 International Conference on Control, Computing, Communication and Materials (ICCCCM), Prayagraj, India, 7 2024, pp. 391-395, doi: 10.1109/ICCCCM61016.2024.11039888.

[45]. Pathak, J., Sakore, N., Kapare, R., Kulkarni, A., & Mali, Y. (2019). Mobile rescue robot. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 4(8), 10-12.

[46]. Hajare, R., Hodage, R., Wangwad, O., Mali, Y., & Bagwan, F. (2021). Data security in cloud. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 8(3), 240-245.

[47]. Y. K. Mali, S. A. Darekar, S. Sopal, M. Kale, V. Kshatriya and A. Palaskar, "Fault Detection of Underwater Cables by Using Robotic Operating System," 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 2023, pp. 10.1109/ICCST59048.2023.10474270.

[48]. Bhongade, A., Dargad, S., Dixit, A., Mali, Y.K., Kumari, B., Shende, A. (2024). Cyber Threats in Social Metaverse and Mitigation Techniques. In: Somani, A.K., Mundra, A., Gupta, R.K., Bhattacharya, S., Mazumdar, A.P. (eds) Smart Systems: Innovations in Computing. SSIC 2023. Smart Innovation, Systems and Technologies, vol 392. Springer, Singapore. https://doi.org/10.1007/978-981-97-3690-4_34

[49]. Y. Mali, M. E. Pawar, A. More, S. Shinde, V. Borate and R. Shirbhate, "Improved Pin Entry Method to Prevent Shoulder Surfing Attacks," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10306875.

[50]. M. Dangore, A. S. R, A. Ghanashyam Chendke, R. Shirbhate, Y. K. Mali and V. Kisan Borate, "Multi-class Investigation of Acute Lymphoblastic Leukemia using Optimized Deep Convolutional Neural Network on Blood Smear Images," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 2024, pp. 1-6, doi: 10.1109/MITADTSoCiCon60330.2024.10575245.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

[51]. A. Chaudhari et al., "Cyber Security Challenges in Social Meta-verse and Mitigation Techniques," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 2024, pp. 1-7, doi: 10.1109/MITADTSoCiCon60330.2024.10575295.

[52]. M. D. Karajgar et al., "Comparison of Machine Learning Models for Identifying Malicious URLs," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-5, doi: 10.1109/ICITEICS61368.2024.10625423.

[53]. Mali, Y.K., Rathod, V.U., Borate, V.K., Chaudhari, A., Waykole, T. (2024). Enhanced Pin Entry Mechanism for ATM Machine by Defending Shoulder Surfing Attacks. In: Roy, N.R., Tanwar, S., Batra, U. (eds) Cyber Security and Digital Forensics. REDCYSEC 2023. Lecture Notes in Networks and Systems, vol 896. Springer, Singapore. https://doi.org/10.1007/978-981-99-9811-1_41

[54]. A. More, S. Khane, D. Jadhav, H. Sahoo and Y. K. Mali, "Auto-shield: Iot based OBD Application for Car Health Monitoring," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-10, doi: 10.1109/ICCCNT61001.2024.10726186.

[55]. J. Pawar, A. A. Bhosle, P. Gupta, H. Mehta Shiyal, V. K. Borate and Y. K. Mali, "Analyzing Acute Lymphoblastic Leukemia Across Multiple Classes Using an Enhanced Deep Convolutional Neural Network on Blood Smear," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/ICITEICS61368.2024.10624915.

[56]. S. Sonawane, U. Mulani, D. S. Gaikwad, A. Gaur, V. K. Borate and Y. K. Mali, "Blockchain and Web3.0 based NFT Marketplace," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724420.

[57]. S. Modi, M. Modi, V. Alone, A. Mohite, V. K. Borate and Y. K. Mali, "Smart shopping trolley Using Arduino UNO," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10725524.

[58]. M. Dangore, D. Bhaturkar, K. M. Bhale, H. M. Jadhav, V. K. Borate and Y. K. Mali, "Applying Random Forest for IoT Systems in Industrial Environments," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725751.

[59]. U. Mehta, S. Chougule, R. Mulla, V. Alone, V. K. Borate and Y. K. Mali, "Instant Messenger Forensic System," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724367.

[60]. Sawardekar, S., Mulla, R., Sonawane, S., Shinde, A., Borate, V., Mali, Y.K. (2025). Application of Modern Tools in Web 3.0 and Blockchain to Innovate Healthcare System. In: Rawat, S., Kumar, A., Raman, A., Kumar, S., Pathak, P. (eds) Proceedings of Third International Conference on Computational Electronics for Wireless Communications. ICCWC 2023. Lecture Notes in Networks and Systems, vol 962. Springer, Singapore. https://doi.org/10.1007/978-981-97-1946-4_2.

[61]. D. R. Naik, V. D. Ghonge, S. M. Thube, A. Khadke, Y. K. Mali and V. K. Borate, "Software-Defined-Storage Performance Testing Using Mininet," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-5, doi: 10.1109/ICITEICS61368.2024.10625153.

[62]. M. Dangore, S. Modi, S. Nalawade, U. Mehta, V. K. Borate and Y. K. Mali, "Revolutionizing Sport Education With AI," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-8, doi: 10.1109/ICCCNT61001.2024.10724009.

[63]. S. P. Patil, S. Y. Zurange, A. A. Shinde, M. M. Jadhav, Y. K. Mali and V. Borate, "Upgrading Energy Productivity in Urban City Through Neural Support Vector Machine Learning for Smart Grids," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-5, doi: 10.1109/ICCCNT61001.2024.10724069.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

[64]. V. Ingale, B. Wankar, K. Jadhav, T. Adedoja, V. K. Borate and Y. K. Mali, "Healthcare is being revolutionized by AI-powered solutions and technological integration for easily accessible and efficient medical care," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10725646.

[65]. Modi, S., Mali, Y., Sharma, L., Khairnar, P., Gaikwad, D.S., Borate, V. (2024). A Protection Approach for Coal Miners Safety Helmet Using IoT. In: Jain, S., Mihindukulasooriya, N., Janev, V., Shimizu, C.M. (eds) Semantic Intelligence. ISIC 2023. Lecture Notes in Electrical Engineering, vol 1258. Springer, Singapore. https://doi.org/10.1007/978-981-97-7356-5_30