iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79309

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

The Crisis of Ground Truth in Medical AI: Evaluating the Tools Used to Detect LLM Hallucinations

Prathamesh Chavan[1], Rushil Dhube[2], Tushar Dayma[3], Riddhi Tumpalliwar[4], Kirti Randhe[5]
ISB&M College Of Engineering, India

Abstract: ***The growing use of Large Language Models (LLMs) in healthcare has produced meaningful advances in diagnostic assistance and clinical documentation. However, the persistent risk of medical hallucinations remains a serious barrier to broader adoption. Detecting fabricated or harmful clinical out- puts requires a reliable foundation of factual correctness, and establishing that foundation in medicine is far more difficult than it first appears. This paper examines what we call the "crisis of ground truth" in medical AI evaluation. We review the tools and methods used to verify AI outputs, organizing the literature around four interconnected themes: the limitations of traditional lexical metrics, the circular reasoning problem in LLM-as-a-judge setups, the challenges of building useful domain-specific benchmarks, and the need to rethink what clinical truth actually means for evaluation purposes. Static benchmarks are highly susceptible to data contamination and struggle to capture multi-turn clinical reasoning. Scalable au- tomated alternatives that use models to judge other models risk validating outputs against themselves rather than against verified medical knowledge. Through thematic analysis of current work, including frameworks such as CLEVER, MedHallBench, and risk-sensitive evaluation methods, we show that automated evaluators can catch obvious factual errors but consistently miss the subtle reasoning failures and safety-critical gaps that clinical environments require. We argue that resolving the ground truth problem requires hybrid evaluation architectures that combine high-throughput automated checks with structured, expert-led human review at key decision points.***
Index Terms: Medical AI, Large Language Models, Halluci- nation Detection, Ground Truth, Review.

I. INTRODUCTION

Large Language Models have demonstrated a remarkable ability to navigate dense medical literature and reason across complex clinical scenarios. Once largely experimental, these systems are now being used to draft patient notes, summarize research, and generate differential diagnoses. Despite these ad- vances, LLMs retain a well-documented tendency to produce confident but factually incorrect statements, a phenomenon commonly referred to as hallucination. In a medical setting, where the cost of error can be severe, even a small divergence from clinical fact can cause patient harm. The medical AI community has responded by developing a range of detection tools and benchmarking frameworks, but all of them share a common underlying problem: to measure whether a model is telling the truth, you need a reliable standard of truth to measure against. In medicine, that standard is much harder to define than it seems [1]–[3].

This is the core of the ground truth crisis. Medical knowledge is vast, context-dependent, and grounded in evolving expert consensus rather than fixed binary facts. For years, evaluation methods relied on static datasets drawn from medical board examinations. These datasets have since been absorbed into the pretraining corpora of modern LLMs, meaning that strong performance on these tests reflects memorization rather than genuine clinical reasoning. This data contamination prob- lem quietly undermines the validity of benchmark scores as indicators of real-world safety [4], [5]. Compounding the issue, measuring accuracy through lexical similarity metrics such as ROUGE or BLEU gives a false sense of security. These metrics reward surface-level word overlap while ignoring the semantic meaning and clinical intent that actually matter [6], [7].

Researchers are pursuing four broad directions to address this problem. The first involves reconsidering how factual accuracy is measured, moving away from lexical proxies toward semantic and risk-calibrated metrics [3], [8]. The second examines the LLM-as-a-judge paradigm, in which large models are used to evaluate other models. While scalable, this approach introduces circular validation and is vulnerable to systematic biases that echo human cognitive weaknesses

[9]–[11]. The third direction addresses domain-specific bench- marking, particularly the shift from single-turn static prompts toward multi-turn, contextually rich evaluation frameworks such as ThreadMedQA and the Swedish Medical LLM Bench- mark [12], [13].

©IJRASET: All Rights are Reserved

The fourth involves a deeper conceptual shift: treating ground truth not as a fixed dataset but as a structural and ontological constraint, drawing on formal logic and topological models to characterize AI reasoning failures [14], [15].

This review provides a structured critique of the tools currently used to detect medical hallucinations, organized around these four themes. Section II describes our literature synthesis process. Section III presents the thematic analysis. Section IV discusses the clinical accountability implications of the ground truth crisis. Section V offers a forward-looking perspective on hybrid evaluation approaches.

TABLE I
COMPARISON OF GROUND TRUTH EVALUATION PARADIGMS



II. METHODOLOGY

To examine the landscape of hallucination detection tools in medical AI and understand how they relate to the ground truth crisis, we conducted a literature review. We searched promi- nent academic databases (including PubMed, IEEE Xplore, and arXiv) using targeted queries such as: *("medical LLM" OR "clinical language model") AND ("benchmark" OR "eval- uation" OR "hallucination detection")*. Through this search strategy, we assembled a corpus of 28 recent academic pub- lications covering medical LLM evaluation strategies, bench- marking frameworks, and hallucination detection methods.

In the first phase of analysis, we focused on the ab- stract, introduction, and methodology of each paper. This approach allowed us to identify the core theoretical claims and evaluation paradigms driving each study without becoming sidetracked by supplementary data or extended bibliographies. In the second phase, we applied a structured qualitative synthesis across the full corpus with two objectives in mind:

1) To extract and document the primary theoretical con- tributions, benchmarking tools, and methodological choices described in each paper.
2) To assess how the empirical findings in each paper either support or complicate the ground truth crisis in medical AI hallucination detection.

This process produced a consistent thematic matrix across the 28 papers. By identifying these recurring themes, we ensured that each contribution was evaluated in relation to our central argument about ground truth validation. The resulting synthesis forms the basis for the thematic analysis in Sec- tion III and supports our four-part classification: the limitations of static benchmarks, the recursive paradox in LLM-as-a-judge setups, the need for fluid domain-specific benchmarks, and the broader epistemological questions about what clinical accuracy means.

III. RESULTS: THEMATIC ANALYSIS

Our literature synthesis points to a significant shift in how the medical AI community thinks about truth and verifi- cation. The field is moving away from static, numerically fixed reference points and toward dynamic, context-sensitive, and structurally enforced evaluation schemas. The tools and frameworks we reviewed cluster into four distinct thematic areas.

A. The Limitations of Lexical Metrics and Static Benchmarks

Lexical metrics such as ROUGE and BLEU were originally developed for general NLP tasks and have long been used as proxies for factual accuracy. Our review surfaces a growing consensus that these metrics are inadequate for clinical appli- cations. Rather than measuring semantic correctness or clinical safety, they operate as surface-level syntactic gauges that can be gamed simply by making outputs longer [6], [7]. A related problem with static benchmarks is that they establish what might be called a rubric-bound ground truth: a model receives a high score because its output matches a narrow predefined checklist, even when it contains clinically dangerous content that falls outside what the metric is designed to detect [16]. As large models are fine-tuned against standardized exams such as MMLU and MedQA, these benchmarks begin to collapse under Goodhart's Law. The metric becomes the target, and its validity as a measure of real-world clinical competence disappears [1], [17].

B. *The LLM-as-a-Judge Paradox*

To avoid the cost and bottleneck of manual expert review, many researchers have turned to the LLM-as-a-judge ap- proach. Tools like PDSQI-9 have shown that advanced models can produce scores that correlate reasonably well with those of human clinicians [18]. The appeal is obvious, but the approach creates a fundamental epistemological problem: a probabilistic system is being used to enforce factual standards over the output of another probabilistic system.

Studies have identified a consistent agreeableness bias in model-based judges: an algorithmic reluctance to reject out- puts that appear plausible on the surface, even when they contain serious factual errors [19]. This means that model judges systematically under-report subtle but clinically signifi- cant fabrications. The problem is compounded by the fact that human evaluators themselves show considerable inter-rater disagreement, so models trained to replicate human judgment end up automating a process that was already imperfect [10], [11].

C. *Domain-Specific and Multi-Turn Benchmarking*

Effective hallucination detection requires tools designed to work beyond single-turn prompts. Clinical diagnosis is an iterative process, and research on multi-turn dialogue has shown that hallucinations tend to cascade: a small factual error in an early exchange can compound into a more serious error as the conversation continues [20]. Evaluation tools that operate on short text fragments are therefore unable to capture the full-context consistency that clinical safety demands.

This points to the need for more nuanced hallucination taxonomies. Frameworks such as the Clinical Safety- Effectiveness Dual-Track Benchmark (CSEDB) make an important distinction: failing to mention a diagnosis (an omission

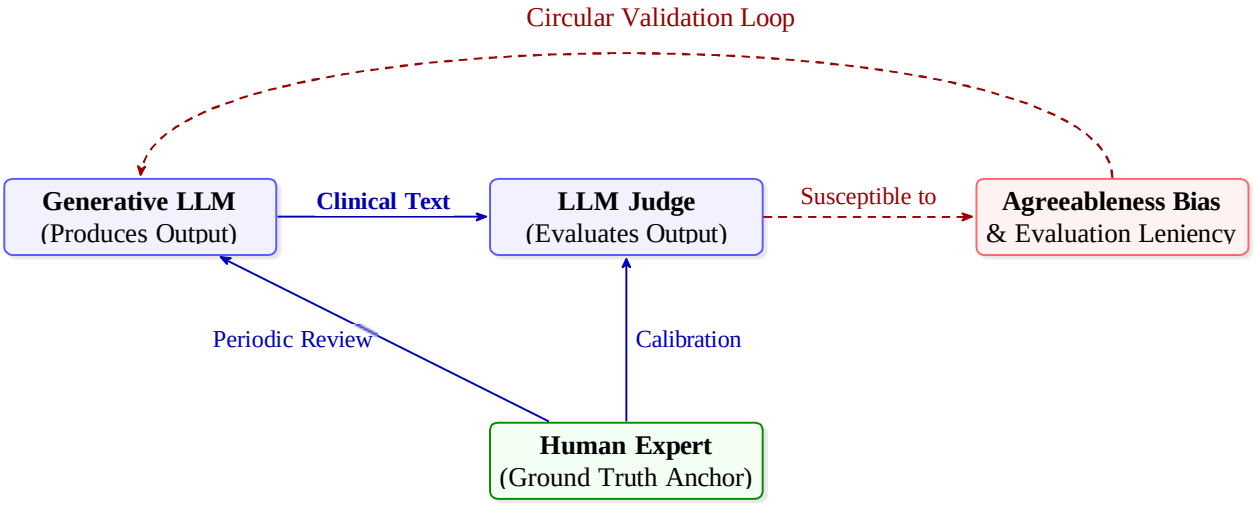


Fig. 1. The recursive truth paradox in LLM-as-a-Judge architectures. Automated judges tend to exhibit agreeableness bias, which inflates true-positive rates while allowing clinically dangerous hallucinations to go undetected. Human expert review anchors the system but cannot realistically scale to cover all model outputs.

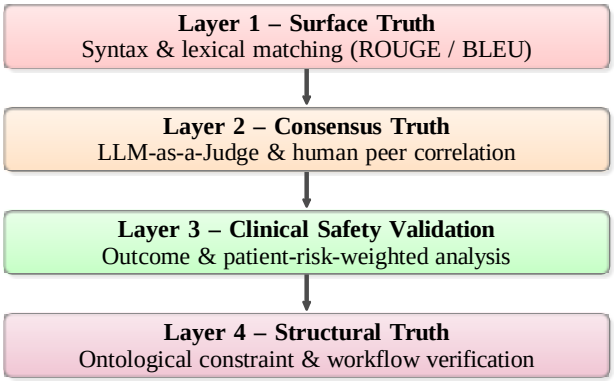


Fig. 2. The layered evolution of ground truth evaluation in medical AI, pro- gressing from surface-level lexical matching toward ontologically grounded structural validation.

or completeness hallucination) can be just as dangerous as fabricating an incorrect one (a consistency hallucination) [17], [21]. Evaluation tools need to be sensitive enough to distin- guish a benign factual slip from a life-threatening fabrication.

D. Epistemological Reframing: Structural and Ontological Validation

Perhaps the most significant conceptual shift visible in recent literature is the move away from treating ground truth as a static dataset and toward treating it as a structural constraint. Drawing on models from adjacent scientific fields, researchers are arguing for physics-informed or ontology-based evaluation mechanisms that test whether a model's reasoning violates established medical logic, such as recommending a pharma- cologically impossible drug-symptom association, rather than simply checking whether output text matches a reference string [4], [17].

In this framing, the ground truth is not a lookup table but a procedural boundary defined by clinical experts. Evaluating a potential hallucination means checking whether the model's output is consistent with that boundary, not simply verifying individual facts against an internet corpus. Expert judgment ul- timately calibrates all automated machinery in this framework [2].

IV. DISCUSSION

Taken together, the reviewed literature makes clear that the ground truth crisis in medical AI is as much a philosophical problem as a technical one. The tools currently used to detect hallucinations are poorly matched to the demands of real clinical practice. As discussed in Section III, relying on lexical overlap or fully automated proxy judges creates serious vulnerabilities: these systems do not account for genuine clinical risk, they display systematic agreeableness bias, and they cannot reliably handle the structural complexity of multi- turn clinical reasoning.

When human experts serve as the anchor for ground truth, scalability breaks down and inter-rater variability introduces noise. When models serve as judges, the evaluation process risks becoming self-referential. The evidence strongly suggests that building trustworthy hallucination detection requires a decisive move toward hybrid architectures. These architectures should combine algorithmic structural validation, checking model outputs against deterministic medical ontologies such as SNOMED-CT, with targeted, risk-weighted human review for complex or ambiguous edge cases.

Accountability is central to this discussion. When a medical LLM produces a hallucinated diagnosis, liability typically falls on the clinician who relied on it. But if the tool designed to catch that hallucination fails because it uses an inadequate definition of ground truth, the systemic risk becomes much harder to manage. Future evaluation frameworks need to treat clinical ground truth not as a fixed binary label but as a risk- stratified, consensus-driven procedural standard.

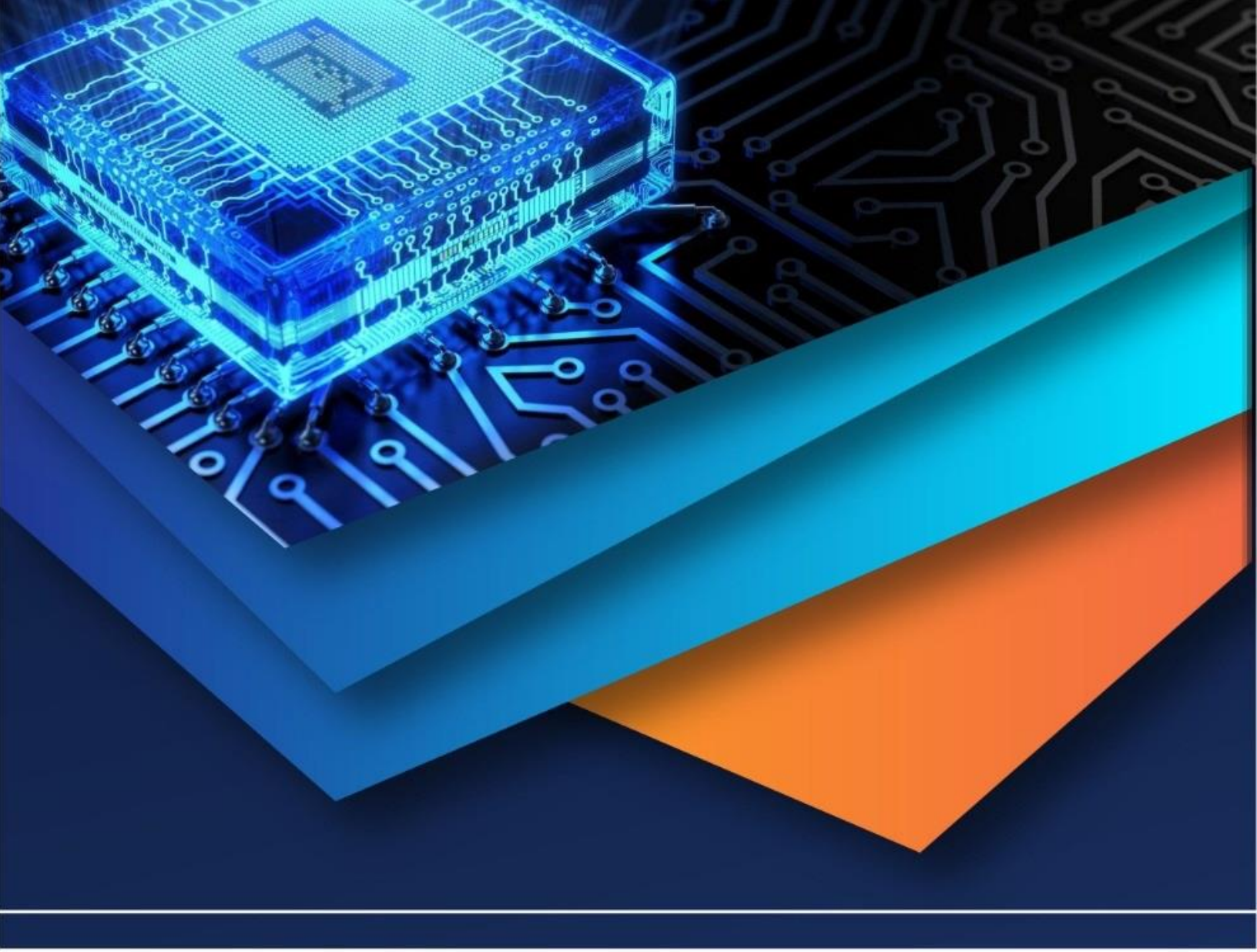
V. CONCLUSION

This review has mapped the current landscape of tools and frameworks being used to address the ground truth crisis in medical AI. We have shown how the field's historical reliance on static benchmarks and lexical metrics falls short when it comes to identifying clinical hallucinations. The LLM-as-a- judge paradigm offers a practical path to scale but introduces evaluation biases that demand careful scrutiny. The most promising direction shifts validation away from isolated fact- checking and toward assessing whether a model's outputs com- ply with the structural logic of constrained medical workflows. Resolving this crisis ultimately requires accepting that ab- solute binary truth in medicine is rarely available. Evaluation systems should instead optimize for clinical safety, support dynamic contextual reasoning, and incorporate meaningful human oversight at the points where it matters most.

REFERENCES

[1] V. Kocaman et al., "Clinical Large Language Model Evaluation by Expert Review (CLEVER): Framework Development and Validation," JMIR AI, 2025. doi: 10.2196/72153
[2] T. Miller et al., "HumanELY: Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization," npj Health Systems, 2025.
[3] S. Doshi, "Beyond Accuracy: Risk-Sensitive Evaluation of Hallucinated Medical Advice," arXiv preprint, 2026. doi: 10.48550/arXiv.2602.07319
[4] S. Gao, J. H. Lau, and J. Qi, "Beyond Seen Data: Improving KBQA Generalization Through Schema-Guided Logical Form Generation," in Proc. EMNLP, 2025. doi: 10.48550/arXiv.2502.12737
[5] S. Pandit et al., "MedHallu-Bench: A benchmark for medical hallucina- tions," arXiv preprint, 2024. doi: 10.48550/arXiv.2412.18947
[6] D. Janiak et al., "The illusion of progress: Re-evaluating hallucination detection in LLMs," Proc. EMNLP, 2025. doi: 10.48550/arXiv.2508.08285
[7] E. Asgari et al., "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," npj Digital Medicine, 8(1), 274, 2025. doi: 10.1038/s41746-025-01670-7
[8] Y. Kim et al., "Medical hallucination in foundation models," Medical Machine Learning, 2025.

[9] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Proc. NeurIPS, 2023. doi: 10.48550/arXiv.2306.05685
[10] R. Williams et al., "Human evaluators vs. LLM judges in clinical decision support," Nature Medicine, 2025.
[11] T. Wang et al., "Bioengineering perspectives on LLMs," Bioengineering, 2025.
[12] Monica Munnangi, Saiph Savage, "ThReadMed-QA: A Multi-Turn Medical Dialogue Benchmark from Real Patient Questions," arXiv preprint, 2026. arXiv:2603.11281
[13] M. Eriksson et al., "The Swedish medical LLM benchmark," Frontiers in Medicine, 2025.
[14] L. Chen et al., "Process vs. outcome in hallucination detection," arXiv preprint, 2025. arXiv:2503.04567
[15] R. Gupta et al., "Semantic illusion in QA models," arXiv preprint, 2025. arXiv:2501.08942
[16] A. Brown et al., "Assessing LLM ability in grading medical notes," JMIR, 2025.
[17] Y. Wang et al., "Clinical Safety-Effectiveness Dual-Track Benchmark (CSEDB)," npj Digital Medicine, 2025.
[18] M. Croxford et al., "Evaluating clinical AI summaries with LLM judges," Health Data Science, 2025.
[19] S. Jain et al., "Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations," arXiv preprint, 2025. arXiv:2510.11822
[20] D. Fan et al., "HalluHard: A Hard Multi-Turn Hallucination Bench- mark," arXiv preprint, 2026. doi: 10.48550/arXiv.2602.01031
[21] K. Zhu et al., "Can We Trust AI Doctors? A Survey of Medical Hallucination in Large Language and Large Vision-Language Models," in Findings of ACL, 2025. doi: 10.18653/v1/2025.findings-acl.350



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)