

A COMPREHENSIVE REVIEW OF LARGE LANGUAGE MODEL EVALUATION, RELIABILITY, AND SAFETY

Gayathri Chilukala

15551 NE Turing Street, #E331, Redmond, WA 98052 United States
gayathrichilukala8422@gmail.com

Received: 15/08/2026

Revised: 25/09/2026

Accepted: 07/10/2026

ABSTRACT:

Large language models (LLMs) have moved rapidly from research artefacts to deployed systems embedded in education, healthcare, software engineering, and autonomous agent pipelines, making rigorous evaluation of their capability, reliability, and safety a matter of direct practical consequence rather than purely academic interest. This paper presents a comprehensive review of the state of LLM evaluation, organised around three interlinked themes. First, it traces the evolution of AI evaluation paradigms, from static, single-task benchmarks of the 1990s and 2000s, through the deep-learning and foundation-model era, to the current agent-centric evaluation challenges posed by autonomous, tool-using LLM systems. Second, it synthesises the principal evaluation methodologies in current use, spanning automatic intrinsic benchmarks, structured human evaluation frameworks scored across dimensions such as relevance, coherence, coverage, and harm, LLM-as-a-judge automatic evaluation, and domain-specific evaluation frameworks illustrated through the case of medical LLM assessment. Third, it examines reliability and safety as distinct but related evaluation concerns, reviewing hallucination and factuality assessment, consistency and robustness testing, inter-rater and human-AI agreement studies, benchmark contamination, alignment techniques including reinforcement learning from human feedback and Constitutional AI, red-teaming methodology, and the emerging challenge of evaluating autonomous LLM-based agents. The review concludes by identifying persistent gaps, including benchmark saturation, the cost and subjectivity of human evaluation, cultural and linguistic bias in existing benchmark suites, and the absence of standardised protocols for long-horizon agentic evaluation, and by outlining practical mitigation strategies, including transparency, shared evaluation infrastructure, and broader stakeholder participation, that can support more trustworthy and reproducible evaluation practice going forward.

Keywords: *Large Language Models, Evaluation, Reliability, AI Safety, Benchmarking, Human Evaluation, Alignment, Red-Teaming, Hallucination, LLM-as-a-Judge*

INTRODUCTION

The capability of large language models has expanded at a pace that has consistently outstripped the maturity of the tools available to evaluate them. Models once assessed primarily on narrow, single-task benchmarks such as sentiment classification or named-entity recognition are now routinely deployed as open-ended conversational assistants, coding copilots, medical information aids, and increasingly as autonomous agents capable of planning, using external tools, and executing multi-step tasks with limited human supervision. Each expansion in capability has been accompanied by a corresponding expansion in the scope of what evaluation must cover: no longer only whether a model produces a correct answer to a closed-form question, but whether its output is factually grounded, internally consistent, appropriately calibrated in its expressed confidence, free of harmful or biased content, resistant to adversarial manipulation, and, in agentic settings, safe and coherent across long horizons of autonomous decision-making.

This breadth creates a genuine evaluation problem rather than a simple measurement problem. Accuracy on a fixed benchmark is a necessary but insufficient signal of real-world reliability, since high benchmark performance can coexist with brittle behaviour under distributional shift, sensitivity to superficial prompt rewording, or outright fabrication of plausible-sounding but false content, commonly termed hallucination. Safety evaluation raises a further set of difficulties: harmful behaviours are often rare, context-dependent, and adversarially elicited, meaning that naive sampling-based evaluation can systematically underestimate risk unless combined with deliberate red-teaming and adversarial probing. Meanwhile, as human evaluation has become both more necessary and more expensive at scale, the field has increasingly turned to LLMs themselves as evaluators, a practice that

introduces its own reliability questions concerning judge bias, self-preference, and agreement with genuine human judgment.

Against this backdrop, this paper undertakes a comprehensive review of LLM evaluation, reliability, and safety, treating these as an integrated evaluation ecosystem rather than three separate literatures. The review is structured to first situate current practice within the historical evolution of AI evaluation paradigms, then to systematically cover the principal families of evaluation methodology in current use, automatic benchmarking, structured human evaluation, LLM-as-a-judge approaches, and domain-specific frameworks, before turning to a focused treatment of reliability (hallucination, consistency, inter-rater agreement, and reproducibility) and safety (alignment methods, red-teaming, harmful-content benchmarks, and agentic safety evaluation) as the two evaluation dimensions of greatest practical consequence for deployed systems. The review concludes with a synthesis of persistent challenges facing the field and a discussion of practical mitigation strategies intended to support more trustworthy, transparent, and reproducible evaluation practice.

OBJECTIVES OF THE REVIEW

This review pursues four specific objectives:

1. To trace the historical evolution of AI and LLM evaluation paradigms, situating current agent-centric evaluation challenges within the broader trajectory from static benchmark-driven assessment to open-ended, real-world evaluation of autonomous systems.
2. To systematically synthesise the principal families of LLM evaluation methodology currently in use, including automatic intrinsic benchmarks, structured human evaluation frameworks, LLM-as-a-judge automatic evaluation, and domain-specific evaluation frameworks, with a focus on their respective strengths, limitations, and appropriate use cases.
3. To critically examine reliability and safety as distinct evaluation dimensions, reviewing hallucination and factuality assessment, consistency and robustness testing, inter-rater and human-AI agreement, alignment techniques, red-teaming methodology, and the emerging challenge of evaluating autonomous LLM-based agents.
4. To identify persistent gaps in current evaluation practice and to outline practical, adoptable mitigation strategies, spanning transparency, shared evaluation infrastructure, cost-sharing, and stakeholder education, that can improve the trustworthiness and reproducibility of LLM evaluation going forward.

EVOLUTION OF AI AND LLM EVALUATION PARADIGMS

Understanding the current state of LLM evaluation requires situating it within the broader historical evolution of artificial intelligence evaluation practice, illustrated in Figure 1. Traditional AI evaluation, spanning roughly the 1990s through the 2010s, was built around static, single-task benchmarks and relied on rule-based and classical machine-learning systems assessed through metrics such as accuracy, precision, and recall on fixed, curated test sets. The deep-learning revolution of the 2012-2020 period, catalysed in large part by large-scale labelled datasets such as ImageNet, shifted the field toward neural network and representation-learning approaches, but evaluation remained largely task-isolated: a model trained and evaluated for image classification was not expected, nor typically assessed, on any other capability.

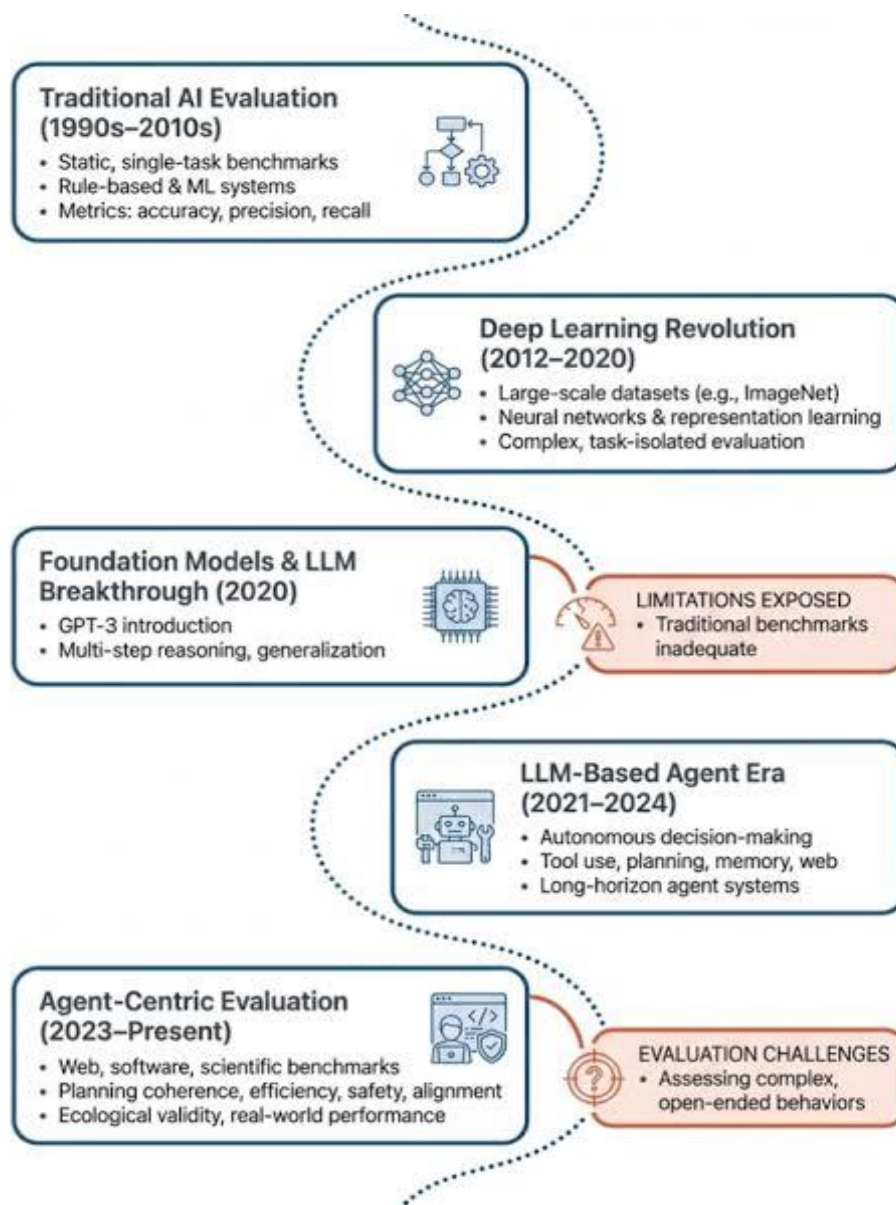


Figure 1. Evolution of AI evaluation paradigms, from traditional static, single-task benchmarking through the deep-learning revolution to foundation-model and LLM-based agent evaluation, culminating in the current agent-centric evaluation era and its associated evaluation challenges.

The introduction of foundation models, most visibly marked by the emergence of GPT-3 in 2020, exposed the inadequacy of traditional, task-isolated benchmarks for models capable of multi-step reasoning and broad generalisation across tasks they were never explicitly trained for. This limitation catalysed the subsequent LLM-based agent era of 2021–2024, characterised by systems capable of autonomous decision-making, tool use, planning, memory management, and long-horizon task execution, capabilities that static single-task benchmarks are structurally unable to assess. The present agent-centric evaluation era, from 2023 onward, has consequently seen the emergence of web, software, and scientific benchmarks specifically designed to assess planning coherence, operational efficiency, safety, and alignment in agentic settings, alongside a growing emphasis on ecological validity, the extent to which benchmark performance predicts genuine real-world performance, as a first-order evaluation concern in its own right rather than an afterthought to accuracy metrics.

This historical trajectory carries a direct methodological implication for the remainder of this review: no single evaluation paradigm inherited from an earlier era is sufficient on its own for contemporary LLMs. Automatic

benchmarking retains value for capability assessment at scale, but must be complemented by human evaluation for open-ended quality judgments, by LLM-as-a-judge methods for scaling that human judgment, and by dedicated safety and reliability testing that earlier, narrower evaluation paradigms were never designed to provide.

EVALUATION METHODOLOGIES

A. Automatic and Intrinsic Benchmarks

Automatic benchmarking remains the most scalable and reproducible evaluation approach available for LLMs, and typically takes one of two forms. Closed-form benchmarks pose questions with a single verifiable correct answer, examples of this class include the Massive Multitask Language Understanding (MMLU) benchmark, which spans 57 academic subjects, GSM8K for grade-school mathematical reasoning, and HumanEval for functional correctness of generated code, all of which permit fully automatic scoring against a ground-truth answer or unit test. Composite evaluation suites such as the Holistic Evaluation of Language Models (HELM) framework and BIG-bench extend this approach by aggregating dozens to hundreds of individual tasks spanning knowledge, reasoning, toxicity, calibration, and efficiency into a single, multi-dimensional evaluation report, explicitly rejecting the premise that any single scalar accuracy figure can characterise a general-purpose model. Open-ended generation tasks, by contrast, resist purely closed-form scoring and have historically relied on reference-based text-similarity metrics such as BLEU and ROUGE, both of which are now widely acknowledged to correlate only weakly with human judgments of quality for genuinely open-ended LLM output, motivating the human and LLM-based evaluation approaches discussed below.

Two further considerations increasingly shape the design of automatic benchmark suites. The first is multilingual and cross-lingual coverage: because the majority of widely cited benchmarks, including MMLU, GSM8K, and TruthfulQA in their original form, were constructed primarily in English, reported capability and safety results have historically offered limited direct evidence about model behaviour in other languages, prompting the development of translated and natively constructed multilingual benchmark variants and, more recently, benchmarks specifically designed to probe cross-lingual consistency, whether a model's answer to the same question changes depending on the language in which it is posed. The second is the growing recognition that raw capability metrics alone are an incomplete basis for model comparison in deployment settings sensitive to cost, and that efficiency-adjusted evaluation, reporting capability per unit of inference cost, latency, or energy consumption, alongside raw accuracy, is necessary for evaluation results to translate meaningfully into deployment decisions, particularly as the range of available models spans several orders of magnitude in both parameter count and inference cost for broadly comparable levels of benchmark performance.

B. Structured Human Evaluation Frameworks

Because automatic metrics remain limited for open-ended generation, structured human evaluation continues to serve as a gold-standard reference against which other evaluation methods are validated. Figure 2 illustrates a representative structured human evaluation pipeline of the kind increasingly used to assess LLM-generated responses: raters, typically recruited through a dedicated web application, score each response across a defined set of factors, relevance (a contextualised composite of accuracy and reasoning quality), coherence (grammatical correctness and overall organisation), comparison (how the response fares against competing entities or alternative model outputs), coverage (whether the response is holistic and complete), and harm (the presence of any harmful component in the response). Per-factor response scores are then aggregated into a final evaluation score per model, permitting systematic, multi-model comparison along each dimension independently rather than collapsing quality into a single undifferentiated rating.

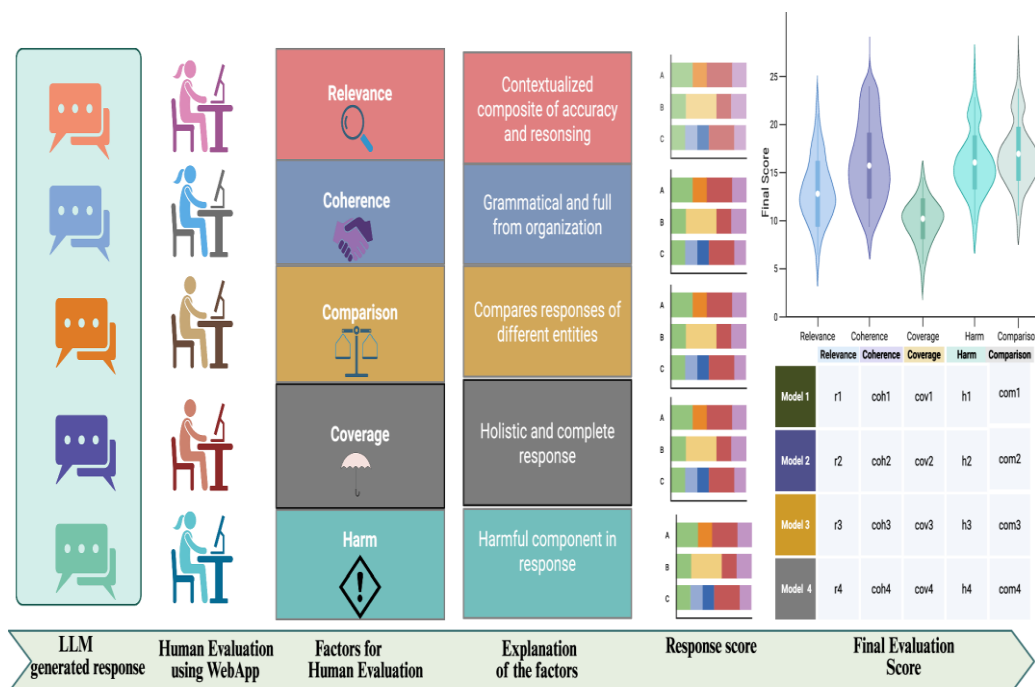


Figure 2. Representative structured human evaluation pipeline for LLM-generated responses, scoring each response across relevance, coherence, comparison, coverage, and harm, with per-factor response scores aggregated into a final evaluation score per model.

This structured, multi-dimensional design addresses a well-documented weakness of unstructured human evaluation, namely that a single undifferentiated quality rating conflates genuinely distinct failure modes (for example, a fluent but factually incomplete response versus an accurate but poorly organised one), making it difficult to diagnose why one model outperforms another or to target improvement efforts. The principal limitations of structured human evaluation remain its cost and throughput at scale, the subjectivity and inter-rater variability inherent in human judgment (examined further in Section V), and the difficulty of recruiting raters with appropriate domain expertise for specialised content such as medical, legal, or technical responses.

C. LLM-as-a-Judge and Automatic Evaluation Using LLMs

To address the scalability limitations of human evaluation, a substantial body of recent work has turned to using LLMs themselves as automatic evaluators, commonly termed the LLM-as-a-judge approach. Methods in this family, including pairwise preference judging as used in the Chatbot Arena and MT-Bench evaluation protocols, and reference-guided scoring approaches such as G-Eval, prompt a strong LLM to compare or score candidate responses according to specified criteria, achieving reported agreement with human preference judgments that is often comparable to inter-human agreement itself on general-purpose conversational tasks. This approach dramatically reduces evaluation cost and turnaround time relative to human panels, and has become a standard component of iterative model development pipelines. However, LLM judges have been shown to exhibit systematic biases, including a preference for longer responses independent of quality, position bias in pairwise comparisons, and a self-preference effect in which a model favours outputs from the same model family as itself, all of which necessitate careful evaluation protocol design, including randomised answer ordering and periodic validation against genuine human judgment, before LLM-as-a-judge results are treated as a reliable substitute for human evaluation.

D. Domain-Specific Evaluation Frameworks: The Case of Medical LLM Evaluation

General-purpose benchmarks are frequently insufficient for high-stakes application domains, motivating the development of domain-specific evaluation frameworks. Figure 3 illustrates a representative framework for evaluating LLMs in the medical domain, structured across four stages: data sources (drawing on both existing curated medical resources and manually curated questions), task scenarios (spanning closed-ended question answering, open-ended dialogue, image processing, and real-world clinical scenarios), evaluation methods (combining automatic evaluation, using algorithms and GPT-based scoring, with human evaluation, drawing on

medical professionals, patients, and computer scientists as distinct rater populations), and finally a set of evaluation dimensions specific to the clinical context, including accuracy, completeness, safety, empathy, reasoning, and hallucination.

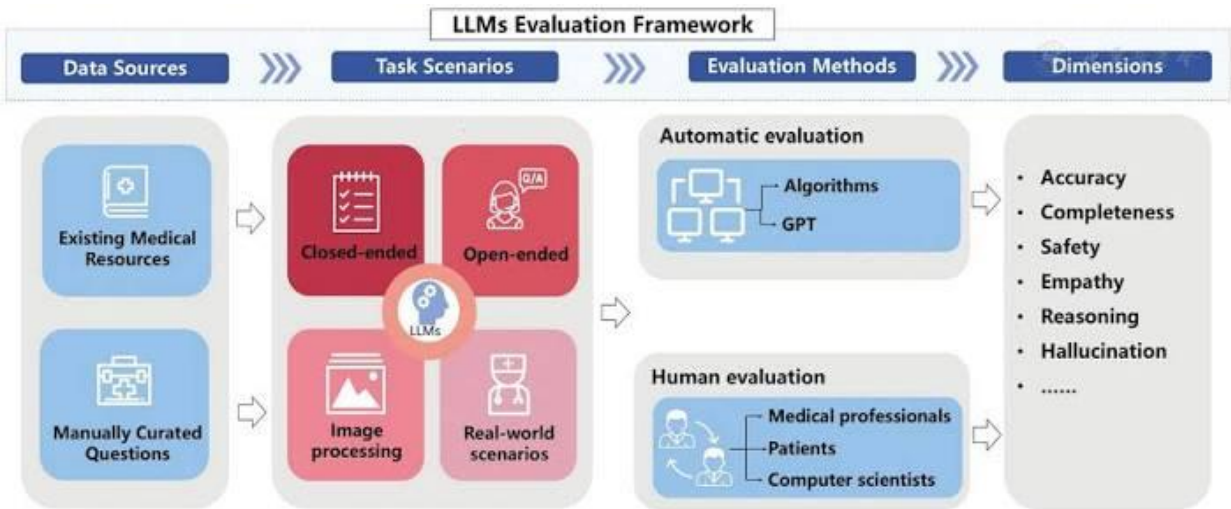


Figure 3. Representative domain-specific LLM evaluation framework for the medical domain, structured across data sources, task scenarios, evaluation methods (automatic and human), and clinically relevant evaluation dimensions including accuracy, completeness, safety, empathy, reasoning, and hallucination.

This domain-specific structure is instructive beyond medicine, since it makes explicit two design principles that generalise to any high-stakes application: first, that evaluation dimensions should be chosen to reflect the specific harms and requirements of the deployment context rather than defaulting to generic accuracy or fluency metrics (empathy and hallucination risk are first-order concerns in a medical dialogue context in a way they are not for, say, a code-completion task); and second, that the rater population itself should be matched to the domain, since general crowdworkers are typically unable to assess clinical accuracy or completeness with the reliability of domain professionals. Comparable domain-specific frameworks have been developed for legal reasoning, financial analysis, and educational applications, in each case adapting both the task scenarios and the evaluation dimensions to the specific risk profile and quality requirements of the domain.

A further point of complexity concerns the growing plurality of applications LLMs are now expected to serve within a single evaluation cycle: the same underlying model may be evaluated as a general-purpose assistant, as a specialised medical or legal information tool, as a coding agent, and as an autonomous web-browsing or tool-using agent, often within the same release cycle. Each of these use cases implies a different evaluation emphasis, general knowledge and instruction-following for the assistant use case, clinical accuracy and empathy for the medical use case, functional correctness for the coding use case, and planning coherence and action safety for the agentic use case, meaning that a single aggregate evaluation score is rarely, if ever, an adequate summary of readiness for deployment. This plurality motivates the structure adopted for the remainder of this review, which treats evaluation methodology, reliability, and safety as related but analytically distinct concerns, each requiring its own dedicated methodology rather than a single unified metric.

Benchmark / Suite	Primary Focus	Scoring Approach
MMLU	Multitask academic knowledge across 57 subjects	Automatic, closed-form multiple choice
HELM	Holistic, multi-dimensional capability and risk profiling	Automatic, aggregated across dozens of scenarios
BIG-bench	Broad, crowd-sourced task diversity	Automatic, closed and semi-open tasks
GSM8K	Grade-school mathematical reasoning	Automatic, exact-match final answer
HumanEval	Functional correctness of generated code	Automatic, unit-test execution

TruthfulQA	Resistance to common human misconceptions	Automatic and human-graded truthfulness
MT-Bench / Chatbot Arena	Open-ended conversational quality	LLM-as-a-judge and crowd pairwise preference
RealToxicityPrompts	Propensity toward toxic continuation	Automatic toxicity classifier scoring

RELIABILITY OF LLMS AND THEIR EVALUATION

A. Hallucination and Factuality

Hallucination, the generation of fluent, plausible-sounding content that is factually incorrect or unsupported by any grounding source, remains one of the most consequential reliability failures in deployed LLM systems, precisely because hallucinated content is often difficult for a non-expert reader to distinguish from genuinely accurate output. Factuality evaluation approaches generally fall into two categories: reference-based methods, which check generated claims against a trusted knowledge source or retrieval corpus, and self-consistency methods, which probe whether a model's answer to the same underlying question remains stable across repeated sampling or paraphrased prompts, on the reasoning that factually grounded knowledge should be more stable under such perturbation than fabricated content. Benchmark suites such as TruthfulQA are specifically constructed around questions that elicit common human misconceptions, testing whether a model reproduces plausible-sounding falsehoods learned from its training distribution rather than genuinely calibrated factual knowledge.

B. Consistency and Robustness

Beyond factuality in the narrow sense, reliability also encompasses a model's consistency and robustness to superficial input variation. A genuinely reliable model should not produce materially different answers to logically equivalent prompts that differ only in surface phrasing, ordering of multiple-choice options, or minor formatting changes, yet such brittleness has been repeatedly documented across model families and remains an active area of adversarial and stress-testing research. Robustness evaluation typically involves systematic prompt perturbation, testing the same underlying query across paraphrased variants, and measuring the resulting variance in model output or accuracy, with high variance interpreted as evidence that benchmark performance may not generalise reliably to the full space of naturally occurring user phrasing.

C. Inter-Rater Reliability and Human-AI Agreement

A further reliability question concerns the evaluation process itself: how consistent are human raters with one another, and how well do automatic or LLM-based evaluators agree with human judgment. Figure 4 presents representative inter-rater correlation data of the kind used to validate evaluation protocols, comparing agreement within an experts-only rater subgroup, within a non-expert subgroup, and between human raters and an automatic LLM-based evaluator. Panel (A) illustrates the pattern commonly observed in such validation studies, correlations tend to be somewhat lower within the experts subgroup than within the non-expert subgroup, reflecting greater diversity of specialist judgment, while agreement between averaged human ratings and automatic evaluation is frequently comparable to, or exceeds, agreement between individual expert and non-expert raters, supporting the use of automatic or LLM-based evaluation as a scalable complement to, rather than wholesale replacement for, human judgment. Panel (B) illustrates a complementary use of such rating data, tracking how average ratings shift as additional context (such as biomarkers, wearable-derived data, or broader situational context) is progressively added to a query, across both human-adaptive and automatic-adaptive evaluation protocols, with expert, non-expert, and LLM-based ratings tracked in parallel to assess whether an automatic evaluator responds to added context in a manner consistent with human judgment.

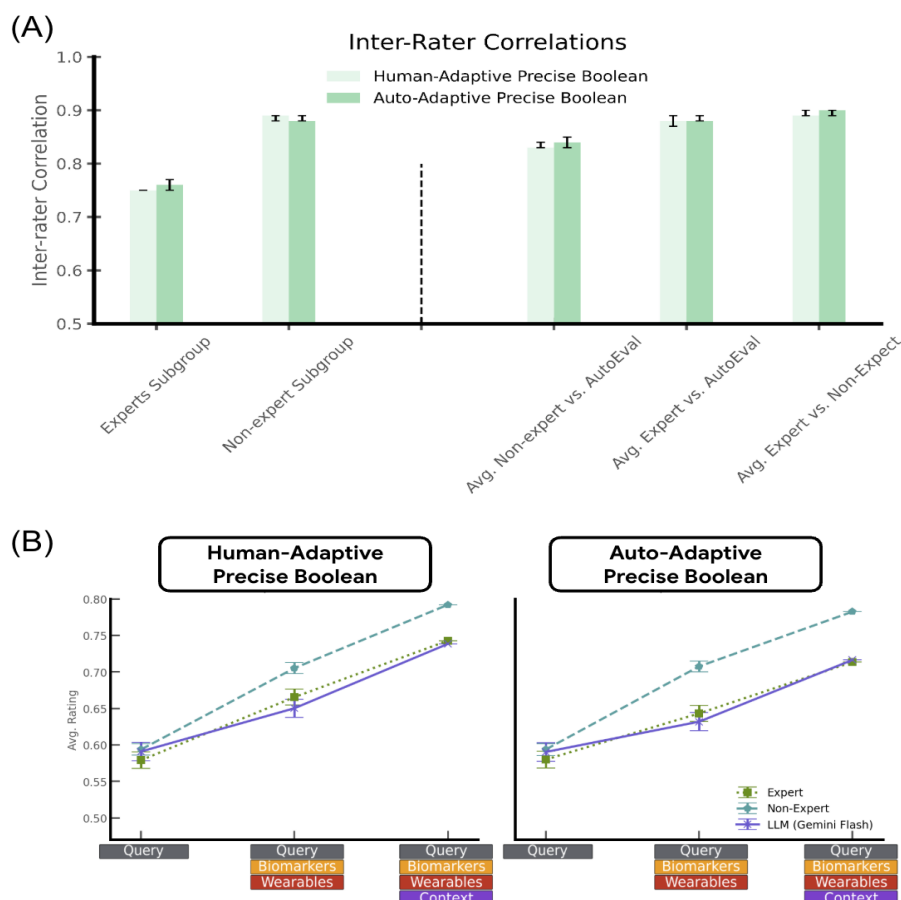


Figure 4. (A) Representative inter-rater correlation analysis comparing agreement within expert and non-expert rater subgroups against agreement between human raters and an automatic LLM-based evaluator. (B) Average ratings as contextual information is progressively added to a query, tracked in parallel across expert, non-expert, and LLM-based raters under human-adaptive and auto-adaptive evaluation protocols.

Such inter-rater reliability analysis serves a validating function for the entire evaluation ecosystem described in Section IV: without demonstrating that human raters agree with one another to an acceptable degree, and that automatic or LLM-based evaluators track human judgment with comparable fidelity, no downstream benchmark result, whether based on human or automatic scoring, can be interpreted with confidence. Reporting inter-rater agreement statistics, commonly using measures such as Cohen's kappa, Krippendorff's alpha, or Pearson and Spearman correlation coefficients, alongside primary evaluation results is accordingly regarded as best practice rather than an optional supplementary analysis.

D. Benchmark Contamination and Reproducibility

A final reliability concern specific to the LLM era is benchmark contamination, the possibility that a model's training corpus has inadvertently or deliberately included benchmark test items themselves, inflating apparent performance without corresponding genuine capability. Because LLMs are trained on web-scale corpora that are difficult to fully audit, contamination detection has become its own sub-field, employing techniques such as n-gram overlap analysis between training data and benchmark items, and canary-string insertion, embedding unique identifying strings into benchmark releases specifically to detect subsequent memorisation. Relatedly, reproducibility of evaluation results has been complicated by the closed or partially closed nature of many widely used commercial models, whose underlying weights, training data, and even inference-time sampling parameters may not be fully disclosed or may change silently between API versions, undermining the assumption of a fixed, stable evaluation target that much of classical benchmarking methodology presupposes.

E. Calibration and Uncertainty Quantification

A related but distinct reliability dimension is calibration, the degree to which a model's expressed or implicit confidence in an answer matches its actual probability of being correct. A well-calibrated model that is uncertain about a claim should ideally signal that uncertainty, whether through explicit hedging language, a numeric confidence estimate, or a refusal to answer, rather than presenting a low-confidence guess with the same fluent assertiveness as a high-confidence fact, since the latter pattern is a principal contributor to user overreliance on hallucinated content. Calibration is commonly assessed using reliability diagrams and expected calibration error computed over a model's token-level output probabilities or over elicited numeric confidence statements, and a consistent finding across the literature is that instruction-tuning and RLHF-style alignment, while improving helpfulness and instruction-following, can measurably degrade calibration relative to the base pretrained model, an effect sometimes termed the alignment tax on calibration and an active area of ongoing mitigation research.

SAFETY OF LARGE LANGUAGE MODELS

A. Alignment Techniques

Contemporary LLM safety practice rests substantially on alignment techniques applied after initial pretraining, intended to shape model behaviour toward helpfulness, honesty, and harmlessness. Reinforcement learning from human feedback (RLHF) trains a reward model on human preference comparisons between candidate outputs, then fine-tunes the base LLM against this learned reward signal using reinforcement learning, an approach popularised through its use in instruction-tuned models such as InstructGPT. Constitutional AI extends this paradigm by using a set of explicit written principles, a constitution, to guide an AI system in critiquing and revising its own outputs, with the resulting preference data then used for reinforcement learning from AI feedback (RLAIF), reducing reliance on large volumes of human-labelled preference data while retaining an explicit, auditable statement of the values the alignment process is intended to instil. Both families of technique remain active areas of methodological refinement, and neither is considered to fully resolve alignment risk on its own, motivating the red-teaming and adversarial evaluation practices discussed next.

B. Red-Teaming and Adversarial Evaluation

Because harmful model behaviours are often rare under naturalistic sampling and may only manifest under deliberately adversarial prompting, red-teaming, the systematic, adversarial probing of a model to elicit harmful, unsafe, or policy-violating outputs, has become a standard component of pre-deployment safety evaluation. Red-teaming methodology spans a spectrum from manual efforts, in which human red-teams craft adversarial prompts (jailbreaks) intended to bypass a model's safety training, to automated approaches in which one LLM is used to generate adversarial prompts against another at scale, substantially increasing coverage of the space of possible attacks relative to manual effort alone. Findings from red-teaming exercises are typically used both to directly patch identified vulnerabilities through additional safety fine-tuning and to construct standing adversarial benchmark suites against which future model versions can be regression-tested, ensuring that previously identified vulnerabilities do not silently re-emerge.

C. Toxicity, Bias, and Harmful-Content Benchmarks

Alongside adversarially elicited harms, LLM safety evaluation must also address harms that can emerge under ordinary, non-adversarial use, principally toxic language generation and demographic or social bias. Benchmark suites such as RealToxicityPrompts assess the propensity of a model to continue a given prompt with toxic language, while bias-focused benchmarks such as ToxiGen and various stereotype-association test suites assess whether model outputs exhibit systematically different quality, sentiment, or content depending on the demographic group referenced or implied in a prompt. A recurring finding across this literature is that safety training measurably reduces, but rarely fully eliminates, both toxicity and bias, and that residual harms are frequently unevenly distributed across languages and cultural contexts, with safety evaluation and mitigation effort to date concentrated disproportionately on English-language and Western cultural contexts relative to the global distribution of actual model deployment and use.

D. Agentic Safety Evaluation

The LLM-based agent era described in Section III introduces safety considerations that extend meaningfully beyond those relevant to single-turn conversational use. An autonomous agent capable of using external tools, executing code, browsing the web, or taking real-world actions on a user's behalf introduces failure modes with potentially direct real-world consequence, including unintended or irreversible actions taken without adequate human oversight, susceptibility to prompt injection from untrusted content encountered during autonomous

operation, and compounding error accumulation across long horizons of sequential decision-making that single-turn evaluation cannot surface. Emerging agentic safety evaluation frameworks accordingly assess planning coherence (whether an agent's sequence of actions remains logically consistent with its stated goal), operational efficiency, and explicit safety and alignment criteria such as whether an agent appropriately declines or seeks human confirmation before taking high-consequence actions, representing the evaluation frontier that Figure 1's historical trajectory identifies as the central open challenge of the current era.

Safety Technique	Mechanism	Principal Limitation
RLHF	Reward model trained on human preference comparisons, used to fine-tune via reinforcement learning	Costly human preference data; reward hacking risk
Constitutional AI / RLAIIF	AI self-critique against written principles, reducing reliance on human labels	Behaviour bounded by quality of the written constitution
Red-teaming (manual)	Human adversaries craft prompts to elicit unsafe outputs	Limited coverage relative to full attack space
Red-teaming (automated)	One LLM generates adversarial prompts against another at scale	May miss attack styles outside the generator's own distribution
Toxicity/bias benchmarking	Standardised prompts scored for toxic or biased continuation	Residual harms unevenly distributed across languages and cultures
Agentic safety evaluation	Assesses planning coherence and appropriate deferral to human oversight	Standardised protocols still immature relative to deployment pace

E. Transparency and Reporting Standards

A cross-cutting safety practice that complements the specific techniques above is structured, standardised disclosure of evaluation methodology and results, most commonly through model cards and system cards, documents accompanying a model release that report intended use cases, known limitations, evaluated benchmark performance, and identified safety risks in a consistent, comparable format. Such reporting standards serve two functions relevant to this review: first, they allow downstream deployers and researchers to assess whether a model's evaluated capabilities and risk profile are appropriate for a given application without needing to independently reproduce the full evaluation suite; and second, they create an auditable record against which subsequent model versions, or independent third-party evaluation, can be compared, supporting the reproducibility concerns raised in Section V-D. Adoption of such standardised reporting remains uneven across the field, with substantial variation in the depth and independence of disclosed safety evaluation between organisations, a gap that several of the mitigation strategies discussed in Section VII are intended to help close.

CHALLENGES AND MITIGATION STRATEGIES

Synthesising the methodologies reviewed above, several practical challenges recur across the LLM evaluation literature, and a corresponding set of mitigation strategies, illustrated schematically in Figure 5, has emerged in response. Cost remains a central constraint: structured human evaluation and adversarial red-teaming are both labour-intensive, motivating solutions built around shared or distributed cost, including collaborative benchmark consortia, cost-sharing across institutions, and the reuse of pre-trained reward and judge models rather than commissioning fresh human evaluation for every model iteration. Understanding gaps among practitioners and end users regarding both model limitations and appropriate evaluation practice motivate dedicated training and structured stakeholder participation, actively encouraging diverse rater and red-team participation rather than relying on a narrow, homogeneous evaluator pool. Privacy concerns arising from human evaluation of potentially sensitive model outputs, particularly in domains such as healthcare, motivate privacy-preserving evaluation protocols and transparent data-handling practices. Finally, correct and informed use of LLMs by evaluators and end users alike, supported by broader public education and awareness, alongside transparency regarding known model limitations, evaluation methodology, and residual safety risks, and the application of modern evaluation technology, including automated contamination detection and LLM-as-a-judge pipelines validated against human judgment as described in Section V-C, together constitute the practical toolkit through which the challenges identified throughout this review can be incrementally addressed.

These mitigation strategies are most effective when applied jointly rather than in isolation. Cost-sharing and pre-trained judge models, for instance, meaningfully reduce the resource burden of frequent evaluation, but their outputs remain trustworthy only to the extent that transparency and periodic human validation, as discussed in Section V-C, are maintained alongside them; conversely, broad stakeholder participation and rater education improve the diversity and quality of human judgment feeding into evaluation, but achieve limited benefit if the resulting data is not accompanied by the transparent reporting practices discussed in Section VI-E. Organisations adopting these strategies therefore face a coordination problem as much as a technical one: sustained improvement in evaluation practice depends on treating cost, participation, privacy, transparency, and technology adoption as a jointly managed programme rather than as independent line items to be addressed piecemeal as resources allow.

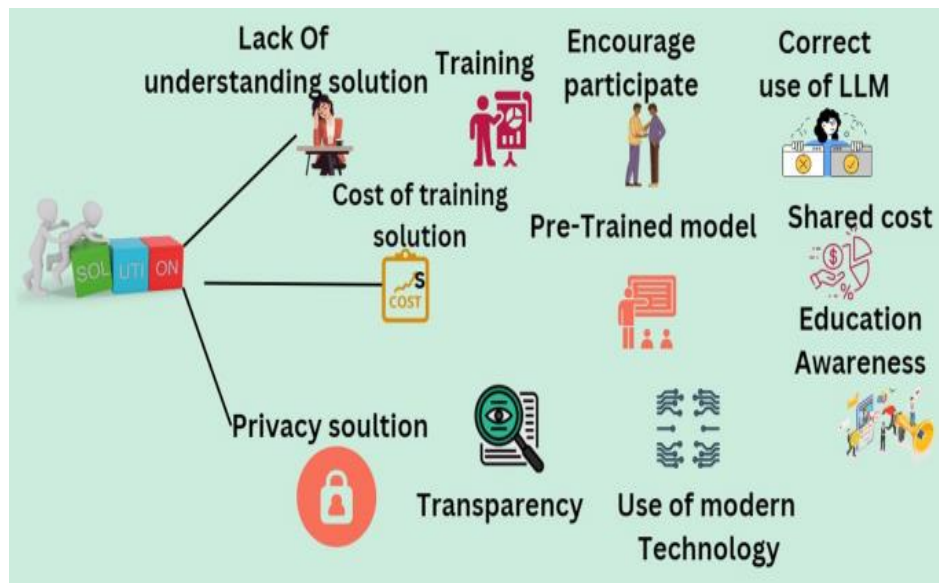


Figure 5. Practical mitigation strategies addressing the principal challenges in LLM evaluation practice: training and stakeholder participation, cost-sharing and use of pre-trained models, privacy-preserving protocols, transparency, correct and informed use of LLMs, and adoption of modern evaluation technology.

DISCUSSION: GAPS AND FUTURE DIRECTIONS

Several gaps persist despite the substantial methodological progress reviewed in this paper. First, benchmark saturation is an increasingly acute problem: as models approach or exceed reported human-level performance on established benchmarks such as MMLU, the discriminative power of these benchmarks to distinguish genuinely stronger models diminishes, motivating continual benchmark refresh but also raising the risk that newer, harder benchmarks are constructed reactively rather than according to a principled theory of the capabilities they are intended to measure. Second, the cost and subjectivity of human evaluation, and the corresponding reliance on LLM-as-a-judge methods, creates a risk of evaluation circularity, in which the models used to judge quality share systematic biases or blind spots with the models being evaluated, a concern only partially addressed by the inter-rater validation practices discussed in Section V-C. Third, cultural and linguistic bias in existing benchmark suites, concentrated disproportionately in English and in Western cultural framings as noted in Section VI-C, limits confidence that reported safety and reliability results generalise to the full global population of actual model users. Fourth, and most acutely for the current agent-centric era identified in Section III, standardised, widely adopted protocols for long-horizon agentic evaluation remain immature relative to the pace at which agentic systems are being deployed, leaving a meaningful gap between demonstrated capability and rigorously validated safety in autonomous, real-world-acting systems.

Addressing these gaps is likely to require continued investment in several directions: dynamic and continually refreshed benchmark suites resistant to saturation and contamination; broader, more linguistically and culturally diverse rater pools and benchmark construction efforts; principled study of LLM-judge bias sufficient to establish clear conditions under which LLM-as-a-judge results can be trusted without independent human validation; and

the maturation of agentic safety evaluation into a standardised discipline comparable in rigour to the single-turn evaluation methodology that preceded it. None of these directions is intractable in principle, but each requires sustained methodological investment commensurate with the pace of capability advancement that motivated this review in the first place.

A further direction meriting explicit attention is the growing interdependence between evaluation and deployment monitoring. Much of the methodology reviewed in this paper, benchmark suites, structured human evaluation, red-teaming, is applied primarily at a fixed point prior to release, yet deployed LLM systems continue to encounter novel prompts, novel user populations, and novel adversarial strategies throughout their operational lifetime, meaning that pre-deployment evaluation alone cannot guarantee sustained reliability or safety once a model is in active use. This has motivated growing interest in continuous or post-deployment evaluation, including live monitoring of user-reported failures, ongoing sampling-based auditing of production traffic for hallucination or harmful content, and periodic re-administration of red-teaming exercises against updated model versions, effectively treating evaluation as an ongoing operational discipline rather than a discrete, one-time gate. Integrating this continuous evaluation perspective with the largely pre-deployment-oriented methodology reviewed in Sections IV through VI represents a natural, and increasingly necessary, extension of current practice. Finally, the review has focused primarily on evaluation methodology developed for and applied to text-based LLMs, yet many of the systems now described as large language models are, in practice, multimodal, accepting image, audio, or video input alongside text. The extension of the evaluation, reliability, and safety methodology reviewed here, benchmark construction, structured human evaluation, hallucination assessment, red-teaming, to genuinely multimodal settings introduces additional methodological complexity, for example, factuality assessment for a claim grounded in an accompanying image requires evaluation protocols distinct from those used for purely text-grounded factuality, that fall largely outside the scope of the present review but represent a natural and important direction for the evaluation research community to pursue as multimodal capability continues to mature alongside the text-based capabilities that have been the primary focus of the literature synthesised here.

CONCLUSION

This review has synthesised the current landscape of large language model evaluation, reliability, and safety, situating contemporary practice within the historical evolution of AI evaluation paradigms and systematically reviewing the principal methodological families in current use: automatic intrinsic benchmarking, structured human evaluation, LLM-as-a-judge automatic evaluation, and domain-specific evaluation frameworks. Reliability and safety were examined as distinct but interdependent evaluation concerns, spanning hallucination and factuality, consistency and robustness, inter-rater and human-AI agreement, benchmark contamination, alignment techniques, red-teaming, harmful-content benchmarking, and the emerging challenge of agentic safety evaluation. The review's central conclusion is that no single evaluation paradigm inherited from earlier eras of AI development is sufficient on its own for contemporary LLMs and the increasingly autonomous systems built upon them; robust evaluation practice instead requires the deliberate, validated combination of automatic, human, and LLM-based methods, explicit attention to the reliability of the evaluation process itself, and evaluation frameworks purpose-built for the specific risk profile of each deployment domain. Persistent gaps, including benchmark saturation, evaluation circularity, cultural and linguistic bias, and immature agentic safety protocols, represent the most pressing directions for future methodological development, and their resolution will be essential to sustaining trustworthy deployment of LLM systems at the scale and consequence now routinely observed in practice.

REFERENCES

1. **Pawan Kumar Singh**, Scaling Gate-All-Around (GAA) Transistor Architectures for Sub-2nm AI Accelerators Using Atomically Thin 2D Materials, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), <https://eudoxuspress.com/index.php/pub/article/view/5713>
2. **Pawan Kumar Singh**, Energy-Efficient Task Scheduling in Green Cloud Computing Using Neuromorphic Spiking Neural Networks, Vol. 33 No. 08 (2024): JOCAAA, Journal Name: Journal of Computational Analysis and Applications, Journal's ISSN: 1521-1398 (Paper),1572-9206 (Online), <https://eudoxuspress.com/index.php/pub/article/view/5712>

3. **Rangavinay Teja Royal Jagga**, 2026, DOI: 10.1109/GCWCN66157.2025.11448495, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,<https://ieeexplore.ieee.org/document/11448495>
4. **Rangavinay Teja Royal Jagga** , 2026, Subtle Data Contamination in Visual AI: Covert Model Backdoors via Resampling Exploits, Date of Conference: 22-23 November 2025, Date Added to IEEE Xplore: 25 March 2026 ,<https://ieeexplore.ieee.org/document/11448514>
5. **Rangavinay Teja Royal Jagga**, 2026, Robotic Tactile-Vision Integration for Interactive Object Disentanglement, <https://ieeexplore.ieee.org/document/11448337>
6. **Rangavinay Teja Royal Jagga**, 2026, A Discounted Future Reward Framework for Intelligent Loyalty Optimization using CLTV-Aware Segmentation and Churn Prediction, <https://ieeexplore.ieee.org/document/11497560>
7. **Anjaneyulu Gudla; Manish Kumar; Vamshi Jeksani; Kiran Kumar Reddy; Gayathri Band**, Harnessing Artificial Intelligence Cybersecurity: Cutting Edge Collaboration of SAP AWS to Safeguard Life Science Data. 2026 <https://doi.org/10.1109/aiei69164.2026.11497833>
8. **G. Mahender, Department of CSE(Data Science), Vardhaman College of Engineering ; Vamshi Jeksani; Manish Kumar; Koppuravuri Sri Lakshmi Sruthi; Kiran Kumar Reddy**, SAP Cloud System on AWS for Risk Detection and Integration Artificial Intelligence Scalable Cybersecurity 2026 <https://doi.org/10.1109/aiei69164.2026.11497435>
9. **Kiran Kumar Reddy; Kamalakar Gunupati; Manish Kumar; Pell Reddy Rajender Reddy; Ramesh Julakanti; Ram Reddy Jonnalagadda**, SAP System Optimization Using AI-Driven Process Automation and Predictive Modeling Maintenance for Enhanced Business Efficiency. 2025 <https://doi.org/10.1109/computingcon64838.2025.11376613>
10. **Amilineni, M.V. 2025**. Audit-Safe Agentic RAG Framework for Regulated Enterprise Systems: A Trustworthy AI Architecture for SAP, Finance, Healthcare, and Government Workflows. INTERNATIONAL JOURNAL OF ADVANCES IN SIGNAL AND IMAGE SCIENCES. (Jan. 2025), 34–46. DOI: <https://doi.org/10.29284/7fjzk883>
11. **Pawan Kumar Singh**, ACCELERATING LARGE LANGUAGE MODEL TRAINING USING HYBRID QUANTUM-CLASSICAL VARIATIONAL AUTOENCODERS WITH INTEGRATED PHOTONIC QRAM, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/501>, DOI: <https://doi.org/10.46121/pspc.54.3.45>
12. **Pawan Kumar Singh** , RESILIENCE ASSESSMENT OF MODERN POWER GRID INFRASTRUCTURE AGAINST EXTREME WEATHER USING PHYSICS-INFORMED NEURAL NETWORKS, Vol. 54 No. 1 (2026): January-March 2026, Power System Protection and Control, ISSN-1674-3415,<https://pspac.info/index.php/dlbh/article/view/502>, DOI: <https://doi.org/10.46121/pspc.54.1.64>
13. **Pawan Kumar Singh**, MITIGATING THE EXABYTE STORAGE CRISIS: ENHANCING READ/WRITE EFFICIENCY IN 5D OPTICAL GLASS STORAGE USING SILVER NANOPARTICLE DOPING, [Vol. 38 No. 11s \(2025\)](#), International Journal of Applied Mathematics (IJAM) , pISSN 1311-1728, eISSN 1314-8060 ,<https://ijamjournal.org/ijam/publication/index.php/ijam/article/view/2153>
14. **Mr. Pawan Kumar Singh (Author), Miss. Pari Nidhi Singh (Author)** ,Intelligent Bot, ASIN : B0DBGCQY8X,Publisher : Namya Press, Publication date : November 25, 2022,Language : English, Print length : 105 pages,ISBN-10 : 9390124867,ISBN-13 : 978-9390124862,Item Weight : 7.5 ounces, Dimensions : 6.24 x 0.43 x 9.24

inches,https://www.amazon.com/Intelligent-Bot-Pawan-Kumar-Singh/dp/B0DBGCQY8X/ref=sr_1_5?crid=2VAXN5BRN5SMS&dib=eyJ2IjoiMSJ9.v3OCQvHBpy20RIjfmWh0qbOl-RL0IUJIC8Xa0Pjeqilvo2sbZ_suN3LbSCw_OA_ZD_4lgjRLMGIBwZQQ1ehZEfDpt_MJiOUfVuqqz_Mav0I5_mb6pN5DQog0U7SA826n1DQQ3QAtWwopHy4lC5lfhZXttKj9faiyATwhdtDATSSV1-r7FmzM2VQyIA0jxcfq6E6hp9ECfH1fby0li9kUfGpLxMq4GktlzUIJO5ggrhsQ.G1AjQiVqU4HFsnScz_vz-X6NXCNs3EFSn-RI8mcWWGOc&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book&qid=1787815741&sprefix=pawan+kumar+singh%2C+bo%2Caps%2C450&sr=8-5

15. **Mr. Pawan Kumar Singh (Author)**, STORAGE CRISES DUE TO RISE OF AI: A Comprehensive Guide to Managing the Storage Challenges in the Age of Artificial Intelligence, ASIN : B0HFK12ZCT Publisher : EB1 Profile, Accessibility : Learn more, Publication date : August 17, 2026, Edition : 1st Language : English, File size : 1.7 MB, Screen Reader : Supported, Enhanced typesetting : Enabled X-Ray : Not Enabled, Word Wise : Not Enabled , Print length : 261 pages, ISBN-13 : 979-8951479112, Page Flip : Enabled, https://www.amazon.com/STORAGE-CRISES-DUE-RISE-Comprehensive-ebook/dp/B0HFK12ZCT/ref=sr_1_1?crid=9MF6C4WDVJR5&dib=eyJ2IjoiMSJ9.DQ8fesBch7ELmRnildt-YA.NBSCHLmetvnbhkwIpOoj3s0IVsCV8y6-ZCfCP6U1lc&dib_tag=se&keywords=Pawan+Kumar+Singh%2C+book%2C+eb1+Profile&qid=1787816040&sprefix=pawan+kumar+singh%2C+book%2C+eb1+profile%2Caps%2C420&sr=8-1
16. **Parag Bharatkumar Parikh**, FRAMEWORKS FOR MITIGATING SUPPLY CHAIN DISRUPTIONS AND MATERIAL PRICE VOLATILITY IN SUSTAINABLE BUILDING CONSTRUCTION, Vol. 53 No. 2 (2025): April-June 2025 , , Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/447>, DOI: <https://doi.org/10.46121/pspc.53.2.34>
17. **Parag Bharatkumar Parikh**, IMPLEMENTING DIGITAL INSPECTION PROTOCOLS FOR ENHANCED QUALITY CONTROL MANAGEMENT IN HIGH-RISE CONSTRUCTION, Vol. 54 No. 3 (2026): July-September 2026, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/439>, DOI: <https://doi.org/10.46121/pspc.54.3.12>
18. **Parag Bharatkumar Parikh** , EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Power System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/issue/view/30>, DOI: <https://doi.org/10.46121/pspc.54.1.61>
19. **Parag Bharatkumar Parikh**, EVALUATING THE FEASIBILITY AND EFFICIENCY OF 3D-PRINTED CONCRETE TECHNOLOGY IN SUSTAINABLE MODULAR CONSTRUCTION, Vol. 54 No. 1 (2026): January-March 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/440>, DOI: <https://doi.org/10.46121/pspc.54.1.61>
20. **Ronakkumar shah**, REDEFINING BUSINESS PROCESS MANAGEMENT: THE SYNERGY OF AI AND INTELLIGENT AUTOMATION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/448>, DOI: <https://doi.org/10.46121/pspc.54.3.14>
21. **Ronakkumar shah**, OPERATIONAL ACCOUNTABILITY IN ACTION: METRICS AND MECHANISMS FOR SYSTEMIC SUCCESS, Vol. 53 No. 4 (2025): October-December 2025, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/452>, DOI: <https://doi.org/10.46121/pspc.53.4.40>

22. **Ronakkumar shah**, BEYOND THE BLUEPRINT: AUTOMATED WORKFLOW ENFORCEMENT AND POLICY GOVERNANCE, Vol. 53 No. 2 (2025): April-June 2025 , System Protection and Control, ISSN-1674-3415 ,<https://pspac.info/index.php/dlbh/article/view/451>, DOI: <https://doi.org/10.46121/pspc.53.2.35>
23. **Tushita Agarwal; Swaminathan Sethuraman**, Mapping Customer Behavior: Visual Analysis of Online Retail Interactions, Date Added to IEEE Xplore: 17 February 2026, DOI: <https://doi.org/10.1109/IC2NC67409.2025.11376261>
24. **Tushita Agarwal**, Efficient Data Summarization and Comparison via Grid-Based Statistical Proxies for Large-Scale Datasets, **Date Added to IEEE Xplore: 03 December 2025**, DOI: <https://doi.org/10.1109/ICBATS66542.2025.11258564>
25. **Dip Bharatbhai Patel**, Building scalable business intelligence systems in the cloud: A technical approach, International Journal of Science and Research Archive, 2021, 02(01), 212-215., Received on 03 February 2021; revised on 17 April 2021; accepted on 19 April 2021, DOI: <https://doi.org/10.30574/ijrsra.2021.2.1.0047>
26. **Dip Bharatbhai Patel**, University of North America, Virginia, USA, Home / Archives / Vol. 5 No. 02 (2025): Volume 05 Issue 02 / Articles, <https://www.academicpublishers.org/journals/index.php/ijdsml/article/view/5958>
27. **Sai Akshara Vemuri** , KINEMATIC OPTIMISATION OF SOFT ROBOTIC MANIPULATORS FOR MEDICAL INTERVENTION, Vol. 54 No. 3 (2026): July-September 2026, System Protection and Control, ISSN-1674-3415, <https://pspac.info/index.php/dlbh/article/view/450>
28. **Sai Akshara Vemuri**, <https://www.periodicos.ulbra.org/index.php/acta/article/view/528>
29. **Aziz, A., Chandusha, K., Naga Kiran, B., Prasad, M. L. M., Golla, K., Rajeswari, S., & Katta, S. K. R.** (2026). Exploring Machine Learning Applications for Genomic Data Analysis in Personalized Medicine. International Journal of Drug Delivery Technology, 16(38s), 1166-1173. <https://doi.org/10.25258/ijddt.16.38s.127>, <https://www.periodicos.ulbra.org/index.php/acta/article/view/687>
30. **Mohammad Naffizuddin , Chaganti Ashok ,Sowjanya Gunukula , Rvbs Sarma , Bhavana Sujanamulk , Bharani Krishna Takkella , Chukka Ram Sunil ,Varri Sujana , Ashwini Kumar**, Etiological Factors of Orofacial Clefts in a Hospital-Based Population: A Descriptive Analytical Study, Published: December 10, 2025, DOI: 10.7759/cureus.98868, DOI : <https://doi.org/10.7759/cureus.98868>
31. **Manoj Kumar Kagitha**, Automated Security Compliance in Ci/Cd: A Natural Language Processing (NLP) Approach to Detecting Vulnerabilities in Infrastructure-As-Code (Terraform/Bicep), Publication year 2021, Volume 2021, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/970>
32. **Manoj Kumar Kagitha**, Hybrid Quantum-Classical Resource Scheduling: A Proposed Architecture for Next-Generation Hyperscale Data Centers, Publication year 2024, Volume 2024, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/969>
33. **Manoj Kumar Kagitha**, Self-Healing Cloud Infrastructure: A Hybrid Reinforcement Learning Framework for Predicting Microservice Failures in Azure Kubernetes Service (AKS), Publication year 2023, Volume 2023, Issue 12, <https://computerfraudsecurity.com/index.php/journal/article/view/971>

34. **Manoj Kumar Kagitha**, GREENOPS IN THE CLOUD: AN AI-DRIVEN TELEMETRY ANALYSIS MODEL FOR REDUCING CARBON FOOTPRINT IN HIGH-AVAILABILITY CLUSTERS, Vol. 28 No. 6 (2020): JOCAAA, <https://www.eudoxuspress.com/index.php/pub/article/view/5054>
35. Lakshmi Rahul Reddy Mareddy. (2026). Security and Performance Assessment of Encryption Techniques for Cloud Platforms. International Journal of Communication Networks and Information Security (IJCNIS), 18(1), 68–80. Retrieved from <https://ijcnis.org/index.php/ijcnis/article/view/8735>
36. Lakshmi RahulReddy Mareddy, Context-Aware Learning Approaches for Improving Prediction Accuracy in Dynamic Systems, International Journal of Computational Mathematical Ideas (IJCMI), Vol.17 No.3(2025): IJCM Vol17 NO3 2025 <https://www.ijcmi.com/index.php/journal/article/view/15>
37. Lakshmi Rahul Reddy , Mareddy. (2026). Revisiting Autonomy in Artificial Intelligence through Agentic System Design. Journal of Computational Analysis and Applications (JoCAAA), 35(1), 707–713. <https://eudoxuspress.com/index.php/pub/article/view/4780>
38. **Mahendrakumar Kalal**, Lean-Driven SAP Production Planning Optimization Using Machine Learning for Inventory and Throughput Efficiency, Date of Conference: 19-20 March 2026, Date Added to IEEE Xplore: 08 April 2026, ISBN Information:, ISSN Information: DOI: 10.1109/ISDFS69419.2026.11459029, <https://ieeexplore.ieee.org/abstract/document/11459029>
39. **Mahendra kumar Kalal**, SUPPLY CHAIN RESILIENCE THROUGH INTEGRATED SAP PRODUCTION PLANNING – PROCUREMENT SYNCHRONIZATION, Archives / Vol. 54 No. 1 (2026): January-March 2026 / Articles, <https://pspac.info/index.php/dlbh/article/view/259>
40. **Dr Manoj Goel (December 2018)**. THE ATTITUDE OF DIFFERENT STATES STUDENTS TOWARDS PHYSICAL EDUCATION; STUDYING IN PHYSICAL EDUCATION COLLEGES OF MAHARSHTRA STATE, <https://ijeponline.com/index.php/journal/article/view/965/847>
41. **Ilakiya Ulaganathan, Tagore Engg. College, Anna University, Chennai, India, ; FNU Antara; Santhosh Chitraju Gopal Varma; Zokir Mamadiyarov; Ziyadulla Turayev**, Published in: 2025 International Conference on Sustainability, Innovation & Technology (ICSIT), Date of Conference: 22-23 August 2025, Date Added to IEEE Xplore: 19 December 2025, ISBN Information: <https://ieeexplore.ieee.org/document/11294748/references#references>
42. **Shreyas Kasture, Sudhanshu Maurya, Amit Kumar Sharma, Santhosh Chitraju Gopal Varma, Kashish Mirza, Firdous Sadaf Mohammad Ismail**, DOI: <http://doi.org/10.11591/ijai.v15.i2.pp1746-1761>, IAES International Journal of Artificial Intelligence (IJ-AI), ISSN/e-ISSN 2089-4872/2252-8938 <https://ijai.iaescore.com/index.php/IJAI/article/view/26932>
43. **Shiv Shankar Dwivedi, Mr. Santosh Kumar Sharma**, Future-Proofing Data Security with Quantum-Resistant Cryptography, Vol. 11 No. 3 (2025): IJCESN, DOI: <https://doi.org/10.22399/ijcesn.3447>, <https://ijcesn.com/index.php/ijcesn/article/view/3447>
44. **Vijender Kalmotia**, Enhancing statistical accuracy in gene perturbation studies, Volume 266, August 2026, 105819, <https://doi.org/10.1016/j.biosystems.2026.105819>, <https://www.sciencedirect.com/science/article/pii/S0303264726001292>

45. **Naveen Undrathi**, Design and implementation of generative AI-driven data transformation pipelines using spring boot for scalable financial applications, eISSN: 2581-9615 || CODEN: WJARAI || Impact Factor 8.2 || CrossRef DOI, September 2026 (Volume 31, Issue 3), DOI url: <https://doi.org/10.30574/wjarr.2026.30.3.1629>,
46. **47: Naveen Undrathi**, Agentic generative AI-enabled microservices framework for intelligent financial data processing in java-based enterprise systems, September 2026 (Volume 31, Issue 3), eISSN: 2581-9615 || CODEN: WJARAI || Impact Factor 8.2 || <https://doi.org/10.30574/wjarr.2026.30.3.1628>
47. **Akshay Jain**, Comparative Analysis of Linkage Methods in Hierarchical Clustering for Customer Segmentation, Conference paper, First Online: 07 October 2025, https://link.springer.com/chapter/10.1007/978-981-96-7289-9_44
48. **Ismail Hossain**, Department of Computer Science, George Mason University, Fairfax, VA, USA, **Mridul Banik**, Department of Computer Science, Colorado State University, Fort Collins, CO, USA, <https://ieeexplore.ieee.org/abstract/document/11446534>
DOI: 10.1109/FMLDS67896.2025.00106
49. **Ismail Hossain, Mridul Banik , Fahmid Al Farid , Jia Uddin , Hezerul bin Abdul Karim ,** Advancing Large Language Models for Low-Resource Languages: A Systematic Review of Pretraining, Adaptation, and Ethical Challenges, 10.32604/cmes.2026.075507, Date: 2026
<https://www.sciencedirect.com/org/science/article/pii/S152614922600322X>
50. **Ismail Hossain; Mridul Banik**, Bridging Visual and Linguistic Intelligence for Chest X-rays: A Comprehensive Review of ViTs and LLM Synergies, Published in: 2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS) Date of Conference: 02-05 November 2025, Date Added to IEEE Xplore: 25 March 2026, ISBN Information: <https://ieeexplore.ieee.org/abstract/document/11446534>
51. **Rifat Sarker Aoyon¹, Ismail Hossain², M. Abdullah-Al-Wadud³, Jia Uddin⁴**, A Secured and Continuously Developing Methodology for Breast Cancer Image Segmentation via U-Net Based Architecture and Distributed Data Training,
<https://www.sciopen.com/article/10.32604/cmes.2025.060917>
52. **M. B. Vibha, M. Chandrika, Samitha Khaiyum & P. Rakshitha Kiran**, Predicting sedentary behavior in adults using stacked LSTM modeling, published: 27 November 2024
Volume 16, pages 346–355 (2025), <https://link.springer.com/article/10.1007/s13198-024-02622-2>
53. **Bellary Chiterki Anil, Jayasimha Sondekoppa Rajkumar, T. L. Divya, Samitha khaiyum Rakshitha Kiran P., Balakrishnan Ramadoss**, A Radiomics-based Framework for Liver Cancer Analysis using Explainable Artificial Intelligence (XAI) Methods, Vol. 15 No. 3 (2025): June, 2025, <https://etasr.com/index.php/ETASR/article/view/10377>
54. **Jayminkumar Patel, Sanjaykumar patel, Birju Patel, Rajan shah**, Development and Validation of a Stability-Indicating Analytical Method for Quantitative Determination of Selected Anticancer Drugs in Pharmaceutical Dosage Form, Archives / Vol. 35 No. S2 (2026): Apr-jun 2026, <https://welltestingjournal.com/index.php/WT/article/view/288>
55. **Jayminkumar Patel, Sanjaykumar patel, Birju Patel, Rajan shah**, Artificial Intelligence-Guided Optimization of Chromatographic Methods for Pharmaceutical Analytical Development, Archives / Vol. 35 No. S2 (2026): Apr-jun 2026,
<https://welltestingjournal.com/index.php/WT/article/view/294>

56. **Sharoon Sharif, Inikpi Egbunu, Yujian Fu, and Zhigang Xiao. 2026.** Hybrid Multi-Modal Learning for Extreme Long-Tail Fungi Classification. In 2026 ACM Southeast Conference (ACMSE 2026), April 23–25, 2026, Troy, AL, USA. ACM, New York, NY, USA, 38–45. <https://doi.org/10.1145/3746467.3801502>