



Global Journal of Research in
Science and Technology

Journal homepage: <https://gsjournals.com/gjrst/>

ISSN: 2980-4167 (Online)



(RESEARCH ARTICLE)



GENERATIVE AI-ENHANCED IT SERVICE MANAGEMENT FOR CLOUD OPERATIONS IN SAUDI GOVERNMENT ORGANIZATIONS

Mohammad Hassan Qureshi *

Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan.

* Corresponding Author

ORCID Details

Mohammad Hassan Qureshi: <https://orcid.org/0009-0008-5999-1124>

Global Journal of Research in Science and Technology, 2026, 04(03), 009–019

Publication history: Received on 18 August 2026; revised on 24 September 2026; accepted on 26 September 2026

Article DOI: <https://doi.org/10.58175/gjrst.2026.4.3.0087>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Generative artificial intelligence is moving IT service management from retrospective ticket administration towards conversational diagnosis, knowledge synthesis and partially automated remediation. Its use in Saudi government cloud operations is nevertheless a high-consequence socio-technical intervention: operational telemetry is sensitive, service interruption can affect citizens, Arabic and English records are heterogeneous, and plausible but incorrect advice can amplify failure. This narrative review critically synthesises thirty peer-reviewed journal articles published from 2020 to 2025. It examines incident intelligence, retrieval-grounded generation, observability correlation, agentic automation, human oversight and public-sector governance, while interpreting their relevance to Saudi digital-government conditions. Evidence was selected through database searching, quality screening and thematic comparison rather than statistical aggregation. The synthesis finds that discriminative models remain preferable for high-frequency detection and classification, whereas generative models add most value in semantic triage, cross-source explanation, runbook adaptation and knowledge capture. The strongest architecture is therefore layered: sovereign data controls and an event fabric support specialised analytics; retrieval constrains a generative reasoning service; policy gates, simulation and human approval control action. Published accuracy results are not directly transferable because datasets, labels and operational definitions differ, and few studies evaluate bilingual, cross-agency or long-running deployments. A risk-tiered adoption pathway is proposed, linking autonomy to evidence quality, reversibility and service criticality. The review identifies research priorities in Arabic operational language, benchmark design, causal diagnosis, continuous assurance, cost-aware inference and measurable public value. It concludes that trustworthy augmentation, rather than model novelty or maximum autonomy, should define generative AI-enabled service management under Saudi digital transformation.

Keywords: Generative Artificial Intelligence; IT Service Management; Cloud Operations; AIOps; Digital Government; Saudi Arabia; Incident Management; Responsible Automation

1 INTRODUCTION

Saudi government organisations increasingly deliver identity, health, municipal, justice and administrative services through interconnected cloud platforms. Operational teams must consequently preserve availability across hybrid infrastructure, application programming interfaces, software services and shared data facilities while responding to incidents whose symptoms cross organisational boundaries. Conventional IT service management provides useful accountability through incident, problem, change and knowledge processes, but its ticket-centred routines often separate human descriptions from machine telemetry. Alert volume, duplicated events, incomplete records and dispersed runbooks make that separation costly. Intelligent alert and incident management research therefore seeks to correlate signals, predict severity and shorten diagnosis [2,3].

Generative artificial intelligence changes the design space because a language model can translate between operator questions, tickets, logs, configuration records and procedural knowledge. Unlike a classifier that returns a category, it can assemble an explanation, ask for missing evidence or draft a recovery plan. Such flexibility is attractive where service desks work bilingually and inherited systems use inconsistent terminology. Yet linguistic fluency is not operational truth. Generated instructions may contain unsupported causes, insecure commands or obsolete dependencies. Public organisations also face stronger expectations of auditability, information sovereignty, continuity and equitable service than ordinary consumer applications. Public-governance scholarship warns that implementation evidence remains less developed than strategic advocacy [17].

This review aims to determine how generative AI can enhance, without weakening, the governability and resilience of cloud IT service management in Saudi government organisations. Six objectives guide the analysis: to evaluate recent methods for detection, triage, diagnosis and remediation; compare generative, discriminative, retrieval-grounded and rule-based approaches; examine how observability and service-management data should be integrated; assess Saudi public-sector applications and readiness constraints; identify evidential, organisational and governance gaps; and formulate an implementable research agenda for trustworthy adoption. The contribution is a context-sensitive synthesis that treats service operations as a controlled decision system rather than a chatbot deployment. It distinguishes tasks that require low-latency numerical inference from those that benefit from semantic generation, then relates autonomy to operational risk.

2 BACKGROUND AND CURRENT RESEARCH LANDSCAPE

IT service management coordinates the design, delivery, restoration and improvement of technology-enabled services. In cloud operations, its unit of concern is the service outcome, whereas observability platforms collect lower-level metrics, logs, traces, topology and events. The operational gap appears when a CPU anomaly, failed deployment and citizen-reported symptom enter different tools without a shared service context. AIOps applies statistical learning, machine learning and automation to bridge this gap. Recent surveys organise the field around failure perception, prediction, root-cause analysis and remediation [2,3]. This sequence is useful, but real incidents are iterative: new evidence changes hypotheses, actions modify telemetry, and operators revise the incident narrative.

Generative models add a semantic coordination layer. They can summarise a ticket, retrieve similar incidents, interpret diagnostic output, produce a change record and preserve reasoning for later review. Generative AI is best understood as an application embedded within a human-task-technology system, not merely a model endpoint [1]. This distinction matters because value depends on retrieval quality, permissions, workflow integration and operator behaviour. Research on professional use shows that effort is reorganised rather than simply removed; checking, prompting and exception handling become new work [27].

The Saudi setting intensifies both opportunity and constraint. Empirical work on smart-government adoption reports that organisational and environmental conditions influence acceptance alongside perceived technological value [18]. A recent national case study identifies uneven maturity, data-integration difficulty, policy alignment and cross-agency coordination as continuing concerns [20]. Citizen adoption of government chatbots also depends on attitude, expected performance and usability [19]. These findings imply that an operations assistant cannot be evaluated solely through answer accuracy. It must fit Arabic-English practices, agency mandates, cloud-residency controls, procurement arrangements and escalation authority. The relevant research landscape is consequently multidisciplinary: software operations establishes technical capability; information systems explains adoption and work redesign; and digital-government research defines legitimacy, accountability and public value.

3 NARRATIVE REVIEW METHODOLOGY

A narrative review was selected because the topic joins technically heterogeneous studies with organisational and public-governance evidence. The purpose was interpretive comparison and framework development, not pooled effect

estimation. Searches were conducted across Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library and multidisciplinary journal databases. Search strings combined variants of “generative AI”, “large language model”, “AIOps”, “IT service management”, “incident”, “service desk”, “cloud operations”, “observability”, “root cause”, “remediation”, “public sector”, “digital government” and “Saudi Arabia”. Backward and forward citation tracing was used only to locate additional eligible journal articles.

Eligibility was restricted to peer-reviewed journal articles dated 2020–2025, written in English, and directly informing at least one of five domains: service-management automation, operational analytics, generative-model architecture, responsible AI governance or Saudi digital-government adoption. Studies had to describe a method, empirical setting, evaluated artefact or defensible conceptual synthesis. Conference papers, preprints, books, commercial reports, standards, editorials without analytical substance, non-peer-reviewed material and articles outside the date window were excluded. Purely clinical, educational or consumer-chatbot studies were removed unless their governance findings were transferable to public operational systems.

Quality assessment considered clarity of the research question, appropriateness of data and method, transparency of evaluation, validity of comparison, stated limitations and relevance to production operations. For reviews, search transparency and analytical coverage were examined. For experiments, dataset provenance, baselines, metrics, leakage controls and external validity were considered. For conceptual and governance articles, argumentative coherence and linkage to evidence were assessed. No single numerical score was treated as a universal threshold; weaker evidence could inform a limitation but not anchor a design recommendation.

The final corpus comprised thirty articles. A structured extraction sheet recorded task, data source, model family, evaluation method, result, strength, limitation and contextual relevance. Critical thematic synthesis proceeded in three passes: within-theme comparison, cross-theme contradiction analysis, and translation into design propositions. Particular attention was given to mismatches between laboratory accuracy and operational usefulness, between generative fluency and factual reliability, and between enterprise automation claims and public-sector accountability. The resulting themes were detection and prediction, semantic incident work, grounded reasoning, controlled remediation, organisational adoption and continuous assurance. Because this is a narrative review, coverage is analytically purposeful rather than exhaustive; selection judgement and rapid technological change remain limitations.

4 CRITICAL REVIEW OF EXISTING TECHNOLOGIES AND METHODS

Rule engines remain valuable for deterministic thresholds, mandatory controls and known failure signatures. They are inexpensive, explainable and predictable, but deteriorate when topology or workload changes. Supervised classifiers improve ticket categorisation and severity prediction when labels are reliable. A machine-learning help-desk study showed that narrative descriptions and comments materially improved classification over sparse fields [4], while incident-severity research demonstrated strong discrimination on large operational datasets but variable performance across user populations [5]. These results support task-specific models, yet they also expose dependence on local labels, writing practices and class balance.

Unsupervised and self-supervised methods address environments where labelled failures are rare. Autoencoder-LSTM analysis can learn normal cloud-log structure [7], and sentence-transformer comparison can rank anomalies across distinct log types without retraining the encoder [8]. Knowledge-graph embeddings make resource relationships explicit [9], while graph and transformer models capture dependencies among metrics [10, 12, 28, 29]. Reported laboratory performance is often high: one adaptive virtual-machine detector reported accuracy above 96 per cent [11], and a memory-augmented graph transformer achieved a mean F1 score of 95.04 per cent across eight datasets [10]. However, thresholds, anomaly definitions and data curation differ. Accuracy without alert burden, time-to-detect, calibration and recovery impact gives an incomplete operational picture.

Generative models address a different bottleneck: making heterogeneous evidence intelligible. Fine-tuned language models can route service-desk cases and exploit synthetic examples where labelled cases are scarce [6]. Large-model AIOps research extends this capability to incident reports, root-cause narratives and remediation suggestions [2]. The advantage is semantic transfer across records that do not share schemas. The weakness is that next-token prediction does not establish causality. Hallucination research shows that unsupported content can arise from data, training and inference processes [22]. In operations, an eloquent false explanation may be more dangerous than an abstaining classifier because it encourages premature action.

Retrieval-grounded generation reduces this risk by providing authorised runbooks, configuration items, resolved incidents and live telemetry as context. Retrieval is not a guarantee: stale documents, incorrect access control and semantically similar but topologically irrelevant incidents can still mislead. A robust implementation must expose evidence links, retrieval time, configuration version and confidence, then require the model to distinguish observation,

inference and recommendation. Knowledge graphs can constrain retrieval to service dependencies, while smaller models can screen high-frequency signals before expensive generation. This division of labour aligns with evidence that specialised models remain efficient for perception, whereas large models add value to contextual reasoning [2,9].

Agentic systems connect reasoning to tools such as log queries, configuration discovery and orchestration. Intent-driven automation research shows how generative interfaces can translate operator goals into network actions [16], and multimodel lifecycle studies describe increasing AI support across microservice engineering [15]. Yet tool access converts a language error into a state change. An operational agent therefore requires allow-listed functions, least privilege, typed parameters, dry-run simulation, policy-as-code validation, rollback and immutable recording. Autonomous execution is defensible only for reversible, low-impact actions with stable evidence. Production restarts, identity changes, data restoration and security controls should remain approval-gated.

Finally, deployment and lifecycle methods determine whether prototype performance survives organisational change. Multi-cloud management needs unified topology and deployment visibility [14]. Cybersecurity implementation research recommends iterative phases that connect use-case selection, data preparation, evaluation, deployment and monitoring [26]. These principles apply directly to service management: the model, retrieval corpus, prompts, tools and policies form one versioned operational product. Evaluating only the foundation model ignores the main failure surfaces.

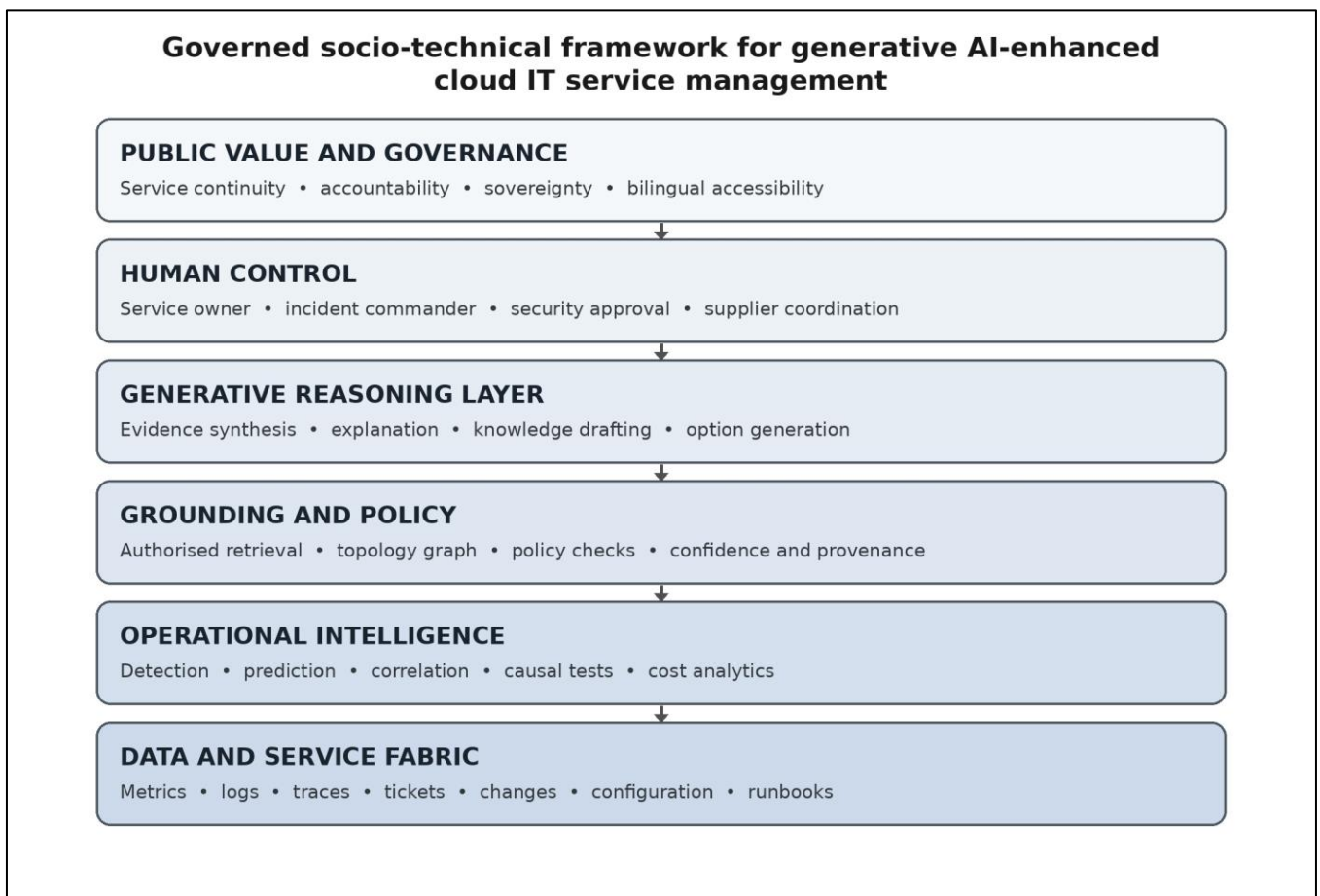


Figure 1: Governed socio-technical framework for generative AI-enhanced cloud IT service management.

5 COMPARATIVE ANALYSIS OF CURRENT APPROACHES

The studies in Table 1 show progress from field-based ticket prediction towards multimodal operational reasoning. They are not directly rankable. Ticket studies optimise category or severity, anomaly studies detect deviations, and generative studies create textual or procedural outputs. Their denominators differ: a correct label may still arrive too late, while a plausible diagnosis may be unactionable. Comparative evaluation should therefore separate perception quality, diagnostic utility, action safety and service outcome.

Table 1: Comparative review of selected recent studies (2020–2025).

Study	Method and setting	Analytical finding	Strength	Limitation and gap
Al-Hawari and Barham [4], 2021	Supervised help-desk classification	Narrative descriptions and comments improve assignment	Uses operational ticket fields	Local labels limit transfer; no generative workflow
Ahmed et al. [5], 2023	Severity prediction on large incident records	Modern language and ensemble models discriminate severity	Multiple populations and models	Performance varies by population; outcome impact untested
Sagbay et al. [6], 2025	Fine-tuned language models with synthetic cases	Domain adaptation supports automatic case assignment	Addresses scarce labels	Synthetic bias and production drift require study
Mitropoulou et al. [9], 2024	Knowledge-graph embedding for cloud anomalies	Topology-aware representation improves contextual detection	Models resource relationships	Operational explanations and action safety absent
Gao et al. [10], 2025	Memory-augmented graph transformer	Robust multivariate anomaly detection across datasets	Broad baseline evaluation	Benchmark performance does not prove service restoration value
Zhang et al. [2], 2025	Survey of large-model AIOps	Models expand data sources, diagnosis and remediation tasks	End-to-end research map	Industrial cost, integration and generalisability remain open
Alkorbi and Alrwais [20], 2025	Saudi digital-government case study	Coordination, integration and skills shape maturity	Direct national relevance	Period and qualitative design limit operational inference

Table 2: Comparative analysis of existing operational approaches.

Approach	Accuracy/reliability	Cost/scalability	Complexity	Readiness	Main advantage	Main limitation
Rules and thresholds	High for known states; brittle under change	Low cost; scales with rule maintenance	Low initially, rising with exceptions	High	Deterministic and auditable	Weak adaptation and event correlation
Supervised classifiers	Strong on stable labelled tasks	Low inference cost; retraining needed	Moderate	High	Fast triage and severity scoring	Label drift and limited explanation
Temporal or graph learning	Strong for nonlinear and dependent signals	Moderate to high compute	High	Medium	Captures dynamic topology and sequences	Opaque thresholds and difficult transfer
Retrieval-grounded generation	Moderate to high when evidence is current	Scalable but inference and indexing costs persist	High	Medium	Inspectable semantic synthesis	Retrieval errors and unsupported inference
Fine-tuned generation	High domain fluency; variable factuality	Training and lifecycle cost high	High	Medium-low	Learns local terminology and formats	Data leakage, drift and costly revalidation

Tool-using agents	Potentially high task completion; low default safety	Scales actions faster than oversight	Very high	Low for critical services	Connects diagnosis with execution	Errors become state changes; strict controls essential
-------------------	--	--------------------------------------	-----------	---------------------------	-----------------------------------	--

Table 2 compares approach families against operational criteria. Rule systems have the highest determinism but the lowest adaptability. Discriminative learning offers mature, economical inference for bounded tasks. Deep temporal and graph models improve representation of dynamic infrastructure at greater training and interpretation cost. Retrieval-grounded generation provides the strongest balance for knowledge-intensive work because evidence can be inspected and updated without retraining. Fine-tuned generation may improve domain language but creates governance and maintenance burdens. Tool-using agents have the greatest potential impact and the lowest default safety.

An important conflict concerns scalability. Cloud-native models are computationally scalable, yet organisational scalability depends on common service identifiers, ownership, knowledge quality and permissions. Similarly, a high benchmark score can coexist with low industrial readiness if the model lacks calibration, drift monitoring or rollback. Readiness should be judged by whether an approach can fail safely under incomplete context, not by accuracy alone. For Saudi government organisations, the preferred portfolio is therefore heterogeneous: lightweight detection at telemetry speed, graph-based correlation at service speed, retrieval-grounded generation at operator speed, and policy-controlled automation at change speed.

6 DECISION RIGHTS AND EVALUATION

Comparison also requires an explicit model of decision rights. A generated summary changes information presentation but leaves authority with the operator; a recommended diagnosis influences judgement; a drafted command creates an executable artefact; and an agent that invokes the command exercises delegated authority. Treating these capabilities as one category obscures their radically different assurance needs. Each use case should be assigned an autonomy level, an accountable owner and a maximum tolerable consequence. Escalation should occur when evidence is missing, sources conflict, confidence falls, the affected service is critical, or the proposed action is irreversible. Abstention is therefore a positive operational outcome rather than a model failure.

Evaluation should mirror this progression. Offline testing can establish retrieval recall, classification calibration, unsupported-claim frequency and command validity. It cannot reveal whether operators interpret uncertainty correctly or whether recommendations shorten restoration. Shadow deployment should replay live incidents without influencing decisions, allowing comparison with actual investigation paths. A subsequent advisory trial can measure acceptance, correction and time saved while preserving human authority. Only bounded actions that repeatedly satisfy prospective safety criteria should enter automated execution. Even then, a control group or phased rollout is needed because concurrent process improvement can otherwise be mistaken for model benefit.

Metrics must represent the whole service system. Technical measures include detection delay, precision, recall, calibration, retrieval coverage, response latency and inference cost. Operational measures include acknowledgement time, restoration time, escalation quality, repeat incidents, failed changes and after-hours workload. Assurance measures include policy violations, unsafe suggestions, unauthorised data exposure, successful rollback and the proportion of outputs with inspectable evidence. Human measures include correction effort, automation bias, skill retention and perceived usefulness. Public-value measures include interrupted transactions, accessibility and service equity. A composite score can conceal unacceptable harm, so safety and privacy should operate as constraints rather than tradable benefits.

This evaluation logic also clarifies procurement. Agencies should test candidate systems against the same de-identified incident packs, tool permissions and latency budgets, while requiring exportable prompts, retrieval indexes, audit records and evaluation results. Contractual service levels should cover model and retrieval degradation, security response, version notification and exit support. Such requirements make portability and institutional learning assessable. They also prevent a polished conversational demonstration from substituting for evidence of controlled performance under realistic operational pressure.

Operational learning should close the loop without converting every correction into training data. A review board should distinguish temporary incident context, verified knowledge and reusable model feedback. Accepted resolutions need service-owner validation, expiry dates and links to the configuration state in which they worked. Rejected suggestions should be coded by failure mode, including missing evidence, wrong dependency, unsafe action or unclear

language. Aggregated patterns can then guide retrieval repair, prompt revision, model selection or process change. This disciplined feedback architecture makes improvement traceable and reduces the risk that yesterday's workaround becomes tomorrow's obsolete instruction.

7 EMERGING TECHNOLOGIES AND DIGITAL TRANSFORMATION

Three developments are likely to reshape operational practice. First, multimodal operational models will connect numerical time series, topology graphs, logs, tickets, screenshots and change records. This can reduce the semantic break between detection and explanation, but only if temporal alignment and service identity are preserved. Second, domain-adapted bilingual models can normalise Arabic and English incident language, technical transliteration and agency-specific abbreviations. Synthetic data may help rare categories, as service-desk research demonstrates [6], but generated examples must be labelled as synthetic and tested for distribution distortion.

Third, operational digital twins can provide a controlled environment for testing generated plans. A proposed action can be simulated against dependency models, capacity constraints and policy before execution. When combined with infrastructure-as-code, the twin also supports counterfactual questions: whether the incident would persist if a deployment were reversed or traffic rerouted. This moves diagnosis from textual resemblance towards causal experimentation.

Digital transformation is not achieved by adding a conversational interface to fragmented processes. Saudi agencies need shared service taxonomies, machine-readable controls and federated knowledge stewardship. Generative AI can then become a boundary object between central cloud teams, agency service owners, security operations and suppliers. Its strategic benefit is improved coordination and institutional learning; its immediate productivity benefit is secondary. Without governance, the same interface can conceal fragmentation behind fluent answers.

8 INDUSTRIAL APPLICATIONS IN SAUDI GOVERNMENT CLOUD OPERATIONS

At the service desk, a retrieval-grounded assistant can clarify requests, identify affected services, translate bilingual descriptions, suggest categorisation and draft responses. The agent should show supporting records and allow staff to correct the proposed classification. Corrections become governed feedback rather than silently entering training data. For major incidents, the assistant can maintain a chronology, correlate changes with symptoms, identify missing diagnostic evidence and prepare stakeholder updates. It should never infer citizen impact solely from alert volume; service-owner confirmation and transaction evidence remain necessary.

In problem management, clustering repeated incidents can reveal hidden demand and technical debt. Generative synthesis can compare clusters with known errors and produce a candidate problem statement, while graph analytics test dependency plausibility. In change enablement, the system can inspect a request for missing test results, affected configuration items, rollback steps and segregation-of-duty conflicts. Code generation may accelerate scripts [25], but validation must occur in a representative environment and approval must remain attributable.

Cloud capacity and cost management offer lower-risk applications. Prediction ensembles can adapt resource reservation to workload drift [13], and an assistant can explain cost anomalies in operational language. The system should distinguish savings from risk transfer: reducing redundant capacity may lower expenditure while weakening resilience. Security operations provide high value but higher consequence. AI can correlate identity, endpoint and cloud events, yet generated containment actions require strict permissions. Public-sector cybersecurity adoption is best treated as an iterative organisational programme rather than a model installation [26].

A phased Saudi deployment should begin with read-only summarisation and knowledge retrieval in one service domain, progress to recommendation after prospective evaluation, and permit bounded execution only after evidence of stable calibration and rollback. Metrics should include mean time to acknowledge and restore, repeat-incident rate, change-failure rate, operator workload, evidence coverage, abstention quality, security events and public-service interruption. This portfolio prevents local speed gains from masking system-wide risk.

Risk-tiered operational workflow from observation to verified learning

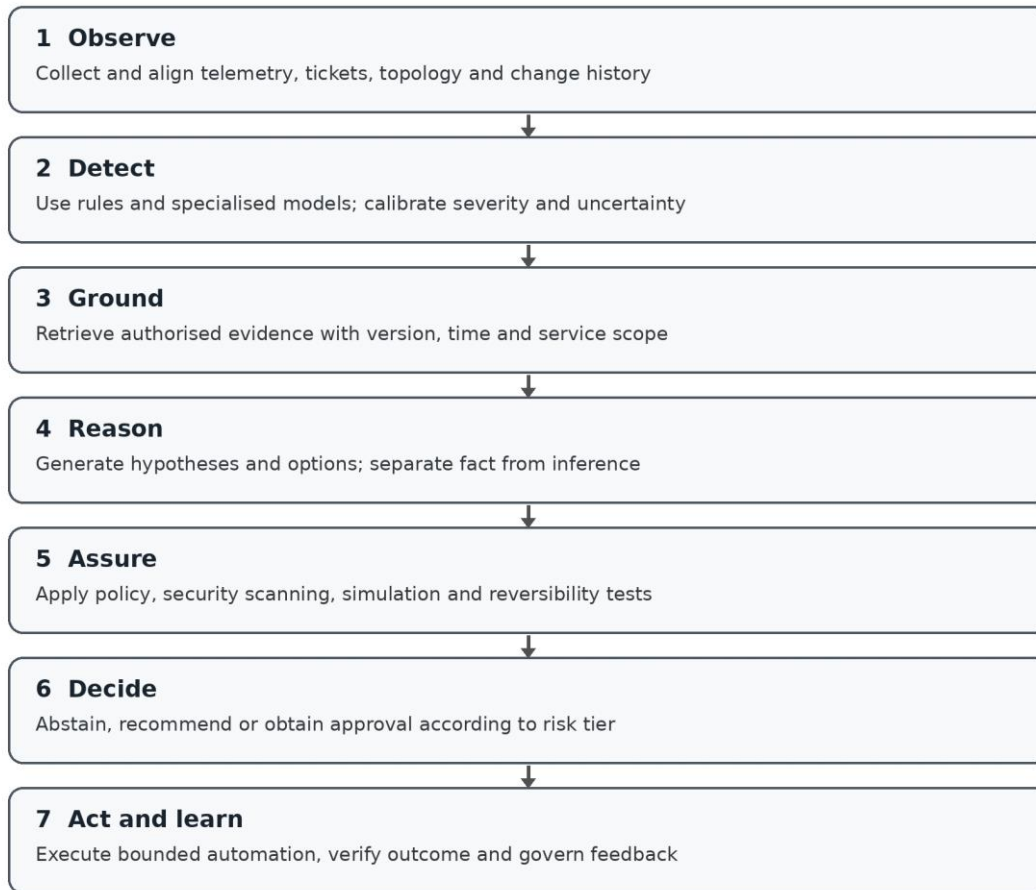


Figure 2: Risk-tiered operational workflow from observation to verified learning.

9 CHALLENGES AND RESEARCH GAPS

The first gap is external validity. Many studies use public datasets or single-enterprise records whose failure distribution, topology and language differ from Saudi multi-agency clouds. Few report prospective deployment, seasonal drift or performance after major architectural change. The second is outcome validity. Classification accuracy and text similarity do not establish faster safe restoration. Evaluation must link model output to operator decisions, recovery quality and unintended consequences.

Data governance is a parallel constraint. Incident records can contain personal information, credentials, vulnerabilities and supplier details. Centralising them for model training can conflict with purpose limitation and agency boundaries. Federated retrieval, redaction, attribute-based access and short-lived context reduce exposure, but each mechanism can remove evidence needed for diagnosis. Research should quantify that privacy-utility trade-off. Arabic operational language is particularly underrepresented: dialect, transliteration, mixed scripts and product names may undermine retrieval even when general language performance appears strong.

Reliability remains unresolved. Hallucination, prompt injection through logs or tickets, malicious runbooks, context-window omission and model drift can corrupt advice [21–24]. Explainability based on a fluent rationale is insufficient because generated rationales may be post-hoc. Evidence provenance, executable tests and counterfactual checks provide stronger assurance. Organisationally, automation can erode operator skills, diffuse accountability and produce automation bias. Public-sector AI research calls for more multidisciplinary and explanatory work [17,30]; this need is acute in operations, where human intervention is required precisely when automation encounters novelty.

Finally, the literature lacks common benchmarks for end-to-end incident handling. Existing comparisons rarely include cost, energy, latency, bilingual performance, security, abstention and reversibility together. There is also limited analysis of vendor exit, model portability and long-term knowledge ownership. These gaps prevent reliable procurement comparisons and risk locking agencies into opaque operational dependencies.

10 FUTURE RESEARCH DIRECTIONS

Future studies should build a privacy-preserving Saudi operational benchmark containing de-identified bilingual tickets, metrics, traces, topology changes and adjudicated incident timelines from multiple service classes. Evaluation should use temporal hold-outs and cross-agency transfer, reporting calibration, false-alert burden, time saved, unsafe recommendation rate and recovery outcome. A shared task could compare specialised models, retrieval-grounded generators and agents under identical evidence and policy constraints.

Causal root-cause analysis is a second priority. Researchers should combine topology, change events and interventions with language models that formulate hypotheses but cannot promote them to causes without tests. Digital-twin experiments can measure whether proposed remediation changes the predicted failure pathway. Third, continuous assurance should treat retrieval corpora, prompts, tools and policies as versioned components. Shadow evaluation, canary release, red-team injection, drift detection and incident replay can reveal regression before broad deployment.

Human-AI collaboration requires longitudinal field research. Studies should examine when operators accept, challenge or over-trust advice; how expertise changes; and which interface makes uncertainty actionable. Bilingual explanations should be evaluated by Saudi operators rather than through automated translation scores. Economic research should compare compact local models, routed model portfolios and remote inference using total cost, latency, energy and sovereignty criteria.

Finally, public value needs explicit measurement. Faster restoration matters because it preserves access to services, not because it improves an isolated technical metric. Future governance models should link every automated capability to a named service owner, risk tier, appeal route and measurable citizen outcome. Cross-agency learning can then occur through federated patterns and control evidence without unrestricted data pooling.

11 CONCLUSION

Generative AI can strengthen cloud IT service management by joining operational evidence, procedural knowledge and human judgement in a usable reasoning interface. Its comparative advantage lies in semantic triage, evidence synthesis, explanation and knowledge capture; high-frequency detection, strict control and routine execution remain better served by specialised analytics and deterministic automation. The review therefore rejects both chatbot-only adoption and unrestricted autonomy.

For Saudi government organisations, a defensible architecture begins with governed data and shared service context, layers specialised detection and graph correlation beneath retrieval-grounded generation, and places policy gates, simulation, approval and rollback before action. This design responds to national conditions of bilingual work, uneven agency maturity, sensitive records and high public-service consequence. The literature provides promising component results but insufficient evidence of longitudinal, bilingual, cross-agency operation. Progress should be judged through safe restoration, calibrated abstention, auditable evidence and public value. A phased, risk-tiered programme can capture practical benefits while generating the empirical evidence that the field currently lacks. Trustworthy augmentation—not maximal model size or autonomy—is the appropriate organising principle for generative AI-enhanced service management.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The author acknowledges the institutional support of Shaheed Zulfikar Ali Bhutto Institute of Science and Technology and contributing scholarly repositories in enabling the synthesis and review of digital government and AIOps frameworks.

Disclosure of conflict of interest

The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

- [1] S. Feuerriegel, J. Hartmann, C. Janiesch and P. Zschech, 'Generative AI', *Business & Information Systems Engineering*, 66, 111–126, 2024. doi:10.1007/s12599-023-00834-7.
- [2] L. Zhang, T. Jia, M. Jia, Y. Wu, A. Liu, Y. Yang, Z. Wu, X. Hu, P. S. Yu and Y. Li, 'A survey of AIOps in the era of large language models', *ACM Computing Surveys*, 58(2), 1–35, 2025. doi:10.1145/3746635.
- [3] Z. Rui, Z. Yu, Z. Shao and X. Yang, 'A survey on intelligent management of alerts and incidents in IT services', *Journal of Network and Computer Applications*, 224, 103842, 2024. doi:10.1016/j.jnca.2024.103842.
- [4] F. Al-Hawari and H. Barham, 'A machine learning based help desk system for IT service management', *Journal of King Saud University – Computer and Information Sciences*, 33(6), 702–718, 2021. doi:10.1016/j.jksuci.2019.04.001.
- [5] S. Ahmed, M. Singh, B. Doherty, E. Ramlan, K. Harkin, M. Bucholc and D. Coyle, 'An empirical analysis of state-of-art classification models in an IT incident severity prediction framework', *Applied Sciences*, 13(6), 3843, 2023. doi:10.3390/app13063843.
- [6] B. Sagbay, M. Zhindón and S. Quevedo, 'Development and fine-tuning of large language models to optimize and automatically assign cases at the service desk', *Intelligenza Artificiale*, 19(1), 66–80, 2025. doi:10.1177/17248035251326656.
- [7] S. Pankajashan, G. Maragatham and G. Devi, 'Hybrid approach with deep auto-encoder and optimized LSTM based deep learning approach to detect anomaly in cloud logs', *Journal of Intelligent & Fuzzy Systems*, 42(6), 6093–6105, 2022. doi:10.3233/JIFS-201707.
- [8] J. Mylläri, T. Aalto and J. K. Nurminen, 'Ladle: a method for unsupervised anomaly detection across log types', *Automated Software Engineering*, 32, 34, 2025. doi:10.1007/s10515-025-00504-w.
- [9] K. Mitropoulou, P. Kokkinos, P. Soumplis and E. Varvarigos, 'Anomaly detection in cloud computing using knowledge graph embedding and machine learning mechanisms', *Journal of Grid Computing*, 22, 6, 2024. doi:10.1007/s10723-023-09727-1.
- [10] H. Gao, R. Xin, P. Chen, X. Li, N. Lu and P. You, 'Memory-augment graph transformer based unsupervised detection model for identifying performance anomalies in highly-dynamic cloud environments', *Journal of Cloud Computing*, 14, 40, 2025. doi:10.1186/s13677-025-00766-5.
- [11] G. Jiang, 'Artificial intelligence-based adaptive anomaly detection technology for IaaS cloud virtual machines', *Journal of Engineering and Applied Science*, 71, 102, 2024. doi:10.1186/s44147-024-00439-1.
- [12] Z. Zhang, Z. Zhu, C. Xu, J. Zhang and S. Xu, 'Towards accurate anomaly detection for cloud system via graph-enhanced contrastive learning', *Complex & Intelligent Systems*, 11, 23, 2025. doi:10.1007/s40747-024-01659-x.
- [13] W. Sus and P. Nawrocki, 'Signature-based adaptive cloud resource usage prediction using machine learning and anomaly detection', *Journal of Grid Computing*, 22, 46, 2024. doi:10.1007/s10723-024-09764-4.
- [14] H. Wei, X. G. Pañeda and J. S. Rodriguez, 'Optimizing machine learning operations in multi-cloud infrastructure: a framework for unified deployment management and topology discovery', *Cluster Computing*, 28, 933, 2025. doi:10.1007/s10586-025-05584-7.
- [15] S. Moreschini, J. Soldani, O. Zimmermann and D. A. Tamburri, 'AI techniques in the microservices life-cycle: a systematic mapping study', *Computing*, 107, 100, 2025. doi:10.1007/s00607-025-01432-z.
- [16] K. Trantzas, A. Kousaridas, N. Baras and P. Demestichas, 'Intent-driven network automation through sustainable multimodal generative AI', *EURASIP Journal on Wireless Communications and Networking*, 2025, 42, 2025. doi:10.1186/s13638-025-02472-x.
- [17] A. Zuiderwijk, Y.-C. Chen and F. Salem, 'Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda', *Government Information Quarterly*, 38(3), 101577, 2021. doi:10.1016/giq.2021.101577.
- [18] S. Alotaibi, 'The determinants of smart government systems adoption by public sector organizations in Saudi Arabia', *Heliyon*, 9(10), e20394, 2023. doi:10.1016/j.heliyon.2023.e20394.
- [19] A. M. Alruwaili, 'Understanding the determinants of using government AI-chatbots by citizens in Saudi Arabia', *International Journal of Electronic Government Research*, 20(1), 1–20, 2024. doi:10.4018/IJEGR.349733.
- [20] S. Alkorbi and O. Alrwais, 'An empirical case study of digital government transformation in Saudi Arabia', *Information*, 16(12), 1110, 2025. doi:10.3390/info16121110.

- [21] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar et al., 'So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI', *International Journal of Information Management*, 71, 102642, 2023. doi:10.1016/j.ijinfomgt.2023.102642.
- [22] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu et al., 'Survey of hallucination in natural language generation', *ACM Computing Surveys*, 55(12), 1–38, 2023. doi:10.1145/3571730.
- [23] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang et al., 'A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions', *ACM Transactions on Information Systems*, 43(2), 1–55, 2025. doi:10.1145/3703155.
- [24] D. Humphreys, A. Koay, D. Desmond and E. Mealy, 'AI hype as a cyber security risk: the moral responsibility of implementing generative AI in business', *AI and Ethics*, 4, 791–804, 2024. doi:10.1007/s43681-024-00443-4.
- [25] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekkki and D. Doermann, 'Future of software development with generative AI', *Automated Software Engineering*, 31, 26, 2024. doi:10.1007/s10515-024-00426-z.
- [26] S. K. Lier, M. Linke, C. Rosenberg and K. B. Brehm, 'An iterative five-phase process model to successfully implement AI for cybersecurity in a corporate environment', *Electronic Markets*, 35, 56, 2025. doi:10.1007/s12525-025-00802-x.
- [27] L. Memmert, D. Soroko and E. Bittner, 'From effort reduction to effort management: an expectancy theory perspective on professionals' work practices with generative AI', *Business & Information Systems Engineering*, 67, 615–635, 2025. doi:10.1007/s12599-025-00960-4.
- [28] A. Alnafessah and G. Casale, 'Artificial neural networks based techniques for anomaly detection in Apache Spark', *Cluster Computing*, 23, 1345–1360, 2020. doi:10.1007/s10586-019-02998-y.
- [29] T. Wang, K. Wang and X. Su, 'Fault localization by analyzing failure propagation with samples in cloud computing environment', *Journal of Cloud Computing*, 9, 17, 2020. doi:10.1186/s13677-020-00164-y.
- [30] V. Rane, 'Advancing smart public administration: challenges and benefits of artificial intelligence', *Urban Governance*, 5, 2025. doi:10.1016/j.ugj.2025.06.003.