International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

AI-Driven IT Service Management in Large Enterprises: Incident Resolution, Service Delivery, and Operational Resilience

Devendra Gairola[1]
[1]*Senior Consultant (Application Development), IBM (San Francisco, CA 94105, USA)*
Email: *devgairola@gmail.com*[1]

Abstract
Enterprise IT has grown too large and too interconnected for manual, rule-based IT service management (ITSM) to keep up. Artificial intelligence for IT operations (AIOps) promises faster detection, routing, diagnosis and recovery, but how much of that holds up in real operations? This structured narrative review examines 12 peer-reviewed primary studies of production AIOps, incident-management and service-desk systems published between 2019 and 2026, alongside four earlier reviews and one industry briefing. It asks how far AI improves incident resolution and service delivery, what makes it work, whether it builds resilience, and how it should be governed. The gains are real but narrow. Ticket routing, diagnosis and outage assessment improve, and two deployments report shorter times to root cause or resolution. Much depends on telemetry quality, dependency knowledge, workflow integration and experienced engineers. Autonomous agents still fail most end-to-end benchmark tasks, and evidence of lasting resilience is thin. Half the primary studies come from one hyperscale provider and none from a regulated sector, so regulated, multi-vendor enterprises should apply the results with care. The paper compares existing AIOps frameworks and proposes a governance-centered framework in which oversight shapes every stage, from data foundations to resilience. A risk-tiered adoption model separates controls backed by the reviewed studies from those drawn from ITSM, security and AI-governance standards, and a research agenda calls for outcome-level, cross-industry evaluation. The central argument is that AI's value in ITSM rests less on model accuracy than on human oversight, grounded knowledge and an organization ready to govern it.

Keywords: *AIOps, IT service management, incident management, large language models, agentic AI, root-cause analysis, service delivery, operational resilience, AI governance*

1. Introduction
Complex, tightly coupled IT systems are essential for large organizations to provide the functionality needed for the delivery of their core business. Operations have become vast and fast with the transition to cloud platforms, microservices, SaaS and hybrid working. It has also resulted in a massive increase in the volume of operational data (logs, metrics, distributed traces, alerts, incident records, service requests and config items). Manual and rule-based analysis are no longer keeping up with the volume of data, and the ops teams are drowning in alerts. Tickets are being routed incorrectly, root causes are taking too long to find, teams are re-troubleshooting the same issues and as a result outages are longer. IT Service Management (ITSM) is a set of practices which organizations utilize to plan, deliver, operate and improve technology-based services [1 - 2]. Incident management is one of the core areas of ITSM, and its aim is to return services to normal and minimize the impact on the user and the business as soon as possible. The old way of doing things would see the service desk and the operations engineers classify and prioritize the incidents, route the tickets to the appropriate resolver group, review the evidence that is available, and then manually determine what to do. This is all very much reliant on the individual's experience and institutional knowledge and can quickly become a massive bottleneck in an organization that has thousands of services and therefore thousands of alerts and requests coming in. AIOps in the context of IT operations is the use of machine learning, big data analytics and automation on operational data in order to support IT management decisions. Remil et al. define AIOps as an emerging paradigm for incident detection and prediction, root cause identification and automated remediation and point out that the field is very much fragmented in terms of terminology, architectures, data and the metrics used to measure performance [3]. Cheng et al. break cloud AIOps down into four functions: incident detection, failure prediction, root

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

cause analysis and automated action [4]. Detection is the question of whether or not a service is behaving abnormally. Prediction is the question of what conditions are likely to lead to a failure in the future. Root cause analysis is about identifying the actual component, dependency, configuration change or event which is the root cause of the failure. Automated action is about the proposal or execution of a response. Those four functions form the basis of moving from a reactive to a predictive and proactive service management model. The high stakes are clearly in the large cloud services where a single production incident can have an impact on millions of users. Ghosh et al. [5] looked at high severity incidents and postmortems from a cloud service that is used by hundreds of millions of people. They looked at the root cause of the incident, the response of the engineers, the process deficiencies that were in place which were preventing recovery and the instances where automation could have been of benefit. The key finding of their study is that incident management can't be based purely on anomaly detection. You need a coordinated response across the entire incident lifecycle from detection and triage through to investigation, mitigation, recovery and post-incident learning. Historical incident data and postmortems are therefore a vital source of knowledge for any AI system that is going to be used to help an organization manage incidents in the future. LLMs have taken AIOps well beyond the stage of classification and prediction. They can now parse unstructured incident descriptions, pull in relevant historical cases, synthesize the diagnostic evidence and propose likely root causes and/or mitigations. Ahmed et al. evaluated LLMs on over 40,000 production incidents across more than 1,000 cloud services, looking at zero-shot, fine-tuned and multitask models and having the incident owners themselves as part of the evaluation [6]. The findings showed that LLMs could help on-call engineers reach root-cause and mitigation conclusions, reducing the time they would otherwise spend poring through documentation and incident history and letting them focus on verifying evidence and selecting actions. Since 2024, research has progressed to tool-using agents that collect telemetry and run diagnostics themselves [7]–[9]. AI can also assist with service delivery even before it reaches a specialist. Service desks are flooded with a huge volume of unstructured requests via email, portal, chat and monitoring tools. Tickets are often misclassified and misrouted between teams, resulting in delayed responses, wasted effort, and SLA breaches. Mandal et al developed a helpdesk ticket dispatch system that uses an ensemble classifier in combination with a configurable rule engine [10]. This has been deployed with three service providers and is currently processing over 90,000 messages per month with around 90% accuracy and coverage. The results are unequivocal in that AI-based classification and routing can deliver operational benefits and can be both flexible and scalable. The benefits are of course accompanied by technical, operational and governance risks. AI-generated suggestions can be wrong, incomplete, out of date or unsupported by the actual evidence. LLMs in particular are capable of generating very convincing (and occasionally quite plausible) justifications which are in fact incorrect. Hallucination and poor predictive accuracy are two of the key problems identified by Zhang et al as being barriers to the use of LLMs for cloud incident root cause analysis [11]. Their approach, LM-PACE (originally released as PACE-LM), uses retrieval augmentation to check that a given prediction is supported by prior incidents and appropriate reference cases. The paper is very clear that confidence calibration, evidence grounding, and human-in-the-loop are all essential when AI is being used to inform operational decisions. Enterprises are up against a whole raft of other issues: poor data quality, not knowing the dependencies between services, a massive array of different monitoring tools, privacy obligations, cyber risk, the problem of legacy integration, model drift and lack of explainability. A model trained on historical incidents is going to re-create those same historical misclassifications and will in fact get worse as the services and infrastructure it's monitoring evolve. The over-reliance on automation can have that well-documented effect of reducing the operator's situational awareness [12] and can allow an incorrect recommendation to propagate through a number of interconnected systems. The right objective for AI in ITSM is therefore not unimpeded autonomy but rather controlled, accountable augmentation with the AI acting as an aide to the operator and with any significant action still being subject to policy, confidence thresholds, audit trails and human oversight [13]. Previous work provides very strong

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

evidence of the individual capabilities of AIOps (anomaly detection, ticket classification, failure prediction, root cause identification, mitigation generation, confidence estimation etc). That work is largely focused on the individual capabilities in isolation. Good benchmark numbers do not, on their own, get service back faster. A dispatch system that routes about 90% of tickets correctly [10] still leaves open whether outages got shorter, whether the same incidents came back, and whether SLIs were met. Answering those questions means looking past the model to the ITSM processes, organizational maturity, governance and resilience goals around it. The earlier reviews are all about the techniques and the tasks. Dang et al. [14] give an industrial perspective on the problems of taking AIOps from the research world into the real world. Notaro et al. [15] develop a taxonomy of failure management based on a review of 100 different approaches. The AIOps studies are categorized by task, data source and technical approach as done by Remil et al. [3] and Cheng et al. [4]. Zhang et al. [16] summarize 183 papers on AIOps based on LLM (up to 2024). These papers say a great deal about what AI can do but comparatively little about organizational readiness, governance or service-level outcomes. Meanwhile, almost all empirical evidence on AI in incident management comes from a small number of hyperscale providers [5]–[8], [11], [17]–[19]. Yet the organizations of interest here—those with formal change control, multi-vendor escalation, legacy platforms, and regulated data—are precisely the ones constrained by these factors. This is true in healthcare, in financial services and in the public sector. The EU is a good example of this. A financial entity in the EU has to comply with the ICT risk management and incident reporting requirements of the Digital Operational Resilience Act [20] and any AI systems used in that environment would be subject to the AI Act [21]. This paper examines the potential of AI-powered ITSM to enhance incident resolution, service delivery and operational resilience in large enterprises. It's based on AIOps reviews and taxonomies, empirical research on production incidents, LLM-based diagnostic systems, deployed service desk automation, confidence estimation systems, recent agentic systems and benchmarks, and the respective ITSM/AI governance standards. It answers four research questions (RQs):

- RQ1. How much does AI actually improve incident resolution and service delivery in large enterprises?
- RQ2. What technical and organizational conditions decide whether it works?
- RQ3. Does AI-enabled incident management strengthen operational resilience?
- RQ4. Which governance mechanisms are needed to manage hallucination, automation risk, explainability, privacy and accountability?

The paper contributes to the answers of these questions as follows:

- C1. As one can see from the different AIOps reviews and taxonomies, all of the existing ones are primarily centered around techniques and tasks, with little to nothing mentioned about governance, organizational readiness or service-level outcomes (see Table 1).
- C2. A lifecycle-level synthesis of the empirical evidence, with each finding linked to the research question it answers (Tables 2 and 3).
- C3. A governance-centered framework in which governance moderates the links between data foundations, AI capabilities, ITSM processes, service outcomes and resilience, rather than acting as a final step before deployment (Section 5 and Figure 2).
- C4. A risk tiered model of adoption that assigns ITSM activities to the controls and assessment metrics, and distinguishes those controls supported by the evidence of the review from those that are proposed, based on the standards; and a research program that prioritizes the regulated multi-vendor enterprise environment (Table 4 and Section 6.4).

Section 2 situates the study within the context of previous AIOps reviews. Section 3 describes the methodology. Section 4 contains the findings and answers to the research questions. Section 5 sets out the proposed framework. Section 6 discusses the study's novelty, the practical implications, the limitations and the directions for future research and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

Section 7 is the conclusion.

2. Related Work and Comparative Positioning

The key references for AIOps research in the ITSM space are four reviews/taxonomies and one industrial briefing. Dang et al. [14] give a very brief industrial perspective on the challenges of moving AIOps from the research stage into production in large online services. Notaro et al. [15] looked at 100 failure management techniques and presented them in a taxonomy of 5 categories and 14 sub-categories according to when the technique operates in the failure timeline and what problem it is addressing. Remil et al. [3] proposed a set of standardized terminology and technical guidelines for AIOps in the context of incident management and classified the literature by task, application area, data source and technical approach. Cheng et al. [4] looked at AIOps in the context of cloud platforms, identified the principal types of telemetry and grouped the tasks as incident detection, failure prediction, root cause analysis and automated action. Zhang et al. [16] reviewed 183 papers on LLM-based AIOps that had been published between 2020 and 2024, looking at the failure data sources, the progression of the tasks, the LLM-based methods, and the evaluation. Table 1 compares these works with the present study on six dimensions, with the characterizations drawn from the abstract and stated scope of each work. There's a very clear pattern. The existing frameworks are all AI-based and heavily focused on task taxonomies and telemetry. They pay very little attention to three questions which are of real significance to the enterprise adopter: how LLM-based assistance should be grounded and governed; what organizational conditions will determine whether technical improvements in fact lead to improvements in service outcomes; and how improvement should be measured at the level of service and resilience rather than at the level of model accuracy. The study in no way supersedes the existing taxonomies. It takes those taxonomies as a basis and adds a governance and outcomes dimension which links the AI capabilities to the changes in ITSM processes and then on to service level and resilience outcomes. It treats governance as a condition that applies at every point in that chain. It draws on a range of sources that the previous reviews don't: ITIL 4 and ISO/IEC 20000-1 for service management practice [1],[2]; the NIST AI Risk Management Framework (including its generative AI profile) and ISO/IEC 42001 for AI governance [22]–[24]; and the resilience engineering research for the concept of operational resilience [25].

3. Research Methodology

3.1. Research Design

This is a narrative review with conceptual synthesis. It is in keeping with the broad guidance of Kitchenham and Charters on literature reviews in software engineering [26] and is not a systematic review or a meta-analysis of the evidence. There is very little evidence on AI in production ITSM, and the range of methodologies is very wide, so the review is based on two types of sources. The primary sources are the actual empirical studies (the studies that have used the AI, the studies that have been looking at AI in actual production systems, the studies that have been looking at AI in the context of real incidents or real world operational environment). These are the sources of the evidence that form the basis of the findings. The secondary sources are the reviews, taxonomies and briefings which are used to inform the analysis and to provide the framework. Standards and regulatory texts are used purely to support the proposed controls in Section 5 and Table 4 and are not used as evidence of the effectiveness of AI. It consists of the following six steps: scoping, searching, screening, extraction and coding, thematic synthesis and the development of a framework (Figure 1).

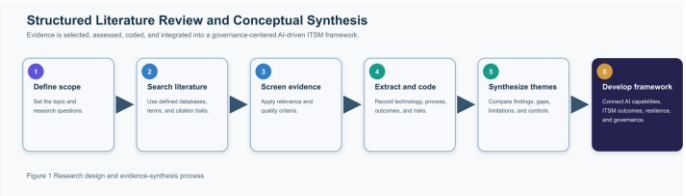

Figure 1 Research design and evidence-synthesis process

3.2. Literature Search Strategy

The search was conducted in the IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, SpringerLink and Google Scholar databases from 2024 to 2026. The search terms were a combination of the concepts of AI and the ITSM processes and enterprise operational outcomes. A common search query is: (“artificial intelligence for IT operations” OR “AIOps” OR “machine learning” OR “large language model” OR “generative AI” OR “intelligent automation”) AND (“IT service

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

management" OR "ITSM" OR "incident management" OR "service desk" OR "help desk" OR "root cause analysis") AND ("service delivery" OR "operational resilience") As shown in Table 1 Comparison of the present study with existing AIOps reviews, taxonomies, and briefings.

Table 1 Comparison of the present study with existing AIOps reviews, taxonomies, and briefings

Work	Type	Lifecycle coverage	LLM / RAG coverage	Governance and oversight	Organizational factors	Outcome focus
Dang et al. [14], 2019	Industrial briefing (2 pages)	AIOps for large-scale online services	Not covered (pre-LLM)	Not a primary focus	Addressed as deployment challenges	Research-to-practice challenges
Notaro et al. [15], 2021	Survey and taxonomy (100 methods)	Failure management across the failure timeline	Not covered (pre-LLM)	Not a primary focus	Limited	Method data requirements and reported results
Cheng et al. [4], 2023 (preprint)	Review of cloud AIOps	Detection, failure prediction, RCA, automated action	Pre-trained language models for log analysis	Not a primary focus	Human-in-the-loop roles noted	Availability; research opportunities
Remil et al. [3], 2024 (preprint)	Literature review and technical guidelines	End-to-end incident-management procedure	Limited	Not a primary focus	Limited	Terminology, taxonomy, and guidelines
Zhang et al. [16], 2025	Survey (183 LLM-based AIOps papers, 2020–2024)	Failure-management tasks and their evolution	Primary focus	Not a primary focus	Not a primary focus	Data sources, task trends, methods, evaluation
This study	**Structured narrative review and conceptual framework**	**Full ITSM lifecycle, incl. service desk and post-incident learning**	**LLM assistance, RAG grounding, confidence calibration, agents**	**Cross-cutting moderator; risk-tiered controls mapped to standards**	**Data foundations, integration, skills, readiness**	**Service-delivery and resilience metrics**

Syntax was database-specific, and searches were restricted to titles, abstracts, and keywords where feasible. The database searches were supplemented by backward citation searching of the most relevant papers and by forward citation searching of key AIOps and incident management papers. The original corpus contained no primary studies published after 2024 so an update search for work published 2024-2026 was performed in October 2026 using the same terms (with 'agent' and 'agentic' added) and with a particular focus on LLM-based and agentic incident management and on AI in service desks in the non-hyperscale provider space. The search window therefore closes in October 2026. These were used to identify candidate sources rather than to generate a complete countable set of results. There were no hit counts or screening numbers recorded, so a PRISMA-style flow diagram is not presented. Rather every source that was included and its role in the synthesis is listed in Section 3.4 and Table 2. This is a limitation and is discussed in Section 6.3.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

3.3. Eligibility Criteria

3.3.1. Inclusion Criteria

- Studies AI, machine learning, deep learning, NLP, generative AI or intelligent automation applied to IT operations or ITSM.
- Covers at least one relevant activity: incident detection, failure prediction, ticket classification, prioritization or assignment, root cause analysis, resolution recommendation, mitigation, or automated remediation.
- Reports or discusses an outcome that matters for this review, whether service delivery, reliability, availability, resilience, governance, or the way people work with AI.
- Any type of contribution qualified: a technical method, an empirical study, a deployed system, a benchmark, a review, a conceptual framework, or an enterprise case study.
- Written in English, with enough bibliographic and methodological detail to assess.
- Primary studies had to be peer-reviewed, in a journal or conference proceedings. Preprints were used only as secondary sources for framing and comparison, never as evidence for the findings.

3.3.2. Exclusion Criteria

- Talks about AI but is entirely irrelevant to IT operations, service management, incident management or enterprise resilience.
- It's all about customer service to the consumer and has nothing whatsoever to do with enterprise IT support.
- Does it lack a clear method, clear evidence, or evaluation?
- Does it include a duplicate, editorial, presentation abstract, news article or non-academic opinion piece?
- Provides no information about the method, data, application or results.
- Not available in full text after reasonable effort.

3.4. Source Selection

Sources were selected deliberately to build up the strongest possible evidence base rather than simply to produce a list of the field. The studies were assessed against the inclusion criteria, and preference was given to peer-reviewed work which either looked at production incident data, reported on systems that had been deployed, or had been evaluated by operational users. The initial search returned 7 primary studies [5], [6], [10], [11], [17]-[19]. The updated search returned 5 more (published between 2024 and 2026) [7]-[9], [27], [28] which were selected mainly to address two gaps in the original set of studies: the dominance of evidence from a single provider and the absence of work on agentic systems. The evidence base is therefore 12 primary studies (see Table 2) and 5 secondary sources: 4 reviews and taxonomies [3], [4], [15], [16] and 1 industrial briefing [14]. 6 of the 12 primary studies were in Microsoft's cloud services, 2 in Alibaba Cloud, 2 from the IBM researchers, 1 in a manufacturing environment and 1 in an academic testbed. The studies differ in terms of the strength of the evidence. The production deployments that report on real-world operational outcomes and the studies which are based on many production incidents are the most convincing. The benchmark and testbed studies [9], [28] are principally used to establish the state of the art with respect to autonomous agents. The implications of this mix are discussed in Section 6.3.

3.5. Quality Appraisal

Each primary study was appraised against the following criteria:

- The study has a very clear research objective.
- The method chosen is appropriate for the aim.
- The operational context and the dataset are adequately described.
- The method used and how it is implemented is sufficiently explained in the context of the AI.
- The means of evaluation and comparison are suitable.
- The evidence reported is in line with the findings.
- Limits and threats to validity are recognized.
- The results are relevant/applicable to enterprise ITSM.

The criteria were applied in a qualitative fashion to determine the relative weight to be given to each

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

study in the synthesis rather than to arrive at some sort of numerical score.

3.6. Data Extraction and Coding

The author extracted the following information for each study and recorded it on a structured extraction form.

- Author, year, title and venue;
- Research design and technique; and
- Enterprise or operational context – provider/organization;
- Dataset type and scale;
- AI or machine-learning technique;
- ITSM process addressed;
- Level of automation;
- Evaluation metrics;
- Principal findings;
- Reported benefits;
- Mistakes, obstacles and pitfalls;
- Governance, security, privacy and human oversight aspects; and
- Help in the resolution of incidents, service delivery or operational stability.

AI methods were coded as conventional machine learning, deep learning, NLP, large language models, retrieval-augmented generation, knowledge-based systems or intelligent/agentic automation. ITSM applications were coded as incident detection, failure prediction, ticket classification, assignment and prioritization, root cause analysis, mitigation recommendation, automated remediation and knowledge management, and post-incident learning.

There are 4 groups of outcome measures. The following measures are considered to be operational efficiency: Mean time to acknowledge, Mean time to resolve, Mean time to recover and Manual effort saved. The measures to be used for service delivery are: Ticket assignment accuracy, First contact resolution, Service level compliance and User satisfaction. The resilience measures are Availability, Downtime, Incident recurrence, Recovery capability and Service continuity. Governance measures are: Model accuracy, Confidence calibration, Explainability, Auditability, Hallucination risk and Frequency of human intervention.

3.7. Evidence Synthesis and Framework Development

The evidence was thematically synthesized rather than subjected to a statistical meta-analysis as the studies are very different in terms of their data sets, AI techniques, operational contexts and evaluation metrics. The aim of the synthesis was to identify the common benefits, enabling conditions, implementation barriers, risks and contradictions. The synthesis is organized around four themes, each tied to one or more research questions. Theme 1 (RQ1) covers what AI can do: detection, prediction, classification, diagnosis, generation, recommendation and automation. Theme 2 (RQ1, RQ2) asks what those capabilities change in ITSM processes, such as faster triage, better assignment, assisted investigation, quicker mitigation and organizational learning, and which technical and organizational conditions help or hold them back. Theme 3 (RQ3) moves up to the enterprise level, where the concerns are efficiency, service quality, reliability, scalability and resilience. Theme 4 (RQ4) is governance, running from data quality and explainability through confidence calibration, security and privacy to human oversight and accountability. More emphasis was placed on peer-reviewed material from large operational datasets, deployed systems, method comparisons, and assessment by operational persons in the interpretation. Findings based just on conceptual arguments or lab work are of course treated with far more skepticism. The last stage was to take the thematic findings and put them together into a governance-centric framework (Section 5). The framework is premised on the enterprise foundations (telemetry quality, knowledge management, system integration, workforce capability and organizational readiness) and connects them with the AI capabilities and ITSM processes. It then connects the process improvements to the results of the services and the resilience of the operations. Governance is not seen as an end goal but rather as a cross cutting enabler. The conceptual evaluation of the framework is done by reviewing the elements of the framework with the patterns, risks and limitations found from the literature. The controls are then mapped to the relevant standards. Those controls which are supported by the literature are clearly distinguished from those which are proposed in the paper.

4. Results and Thematic Findings

The studies reviewed make use of AI in a number of different ways across the ITSM lifecycle in five areas: incident detection and diagnosis, service desk automation, LLM-assisted root cause analysis and mitigation, agentic and tool-using systems, and the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

governance of AI-assisted operational decisions. Table 2 summarizes the 12 key studies. Where two of the studies have the same first author surname they are differentiated by the name of the system in question.

4.1. AI-driven Incident Detection and Diagnosis

Machine learning systems can continuously review logs, performance metrics, alerts, and traces and identify deviations from normal behavior before it becomes a significant disruption for operations teams. Failure prediction models take this a step further by using historical and real-time telemetry to predict future failures. The effectiveness of both approaches depends on the quality, completeness and representativeness of the operational data. However, the detection of the problem does not help to significantly shorten the recovery time unless followed by a correct diagnosis. Alarms can cascade through many dependent applications and infrastructure elements in complex enterprise environments in the event of a single failure. With the help of AI-powered root cause analysis, these alarms can be connected and the different parts can be prioritized based on their probability of being the root cause of the failure. This process significantly restricts the realm of the engineer, but context investigation is still needed. Ghosh et al. [5] show that production incidents go through a sequence of linked phases (detection, root cause analysis and mitigation) and that a delay in any one of those phases will have an impact on the overall recovery process. Therefore AI models shouldn't be deployed in isolation without considering how the output of those models will feed into the subsequent incident management activities. Chen, Z. et al. [17] reach the same conclusion from a two-year study of incident management at Microsoft. They identify two key issues: incomplete service and resource dependencies and imprecise resource health assessment. Both of these they point to as being issues of system design and operations rather than anything to do with the model. Jin et al. [19] note that a key task during a major outage is to determine the scope of impact by aggregating related incidents; their Oasis system does this and also generates LLM-based outage summaries. BiAn [8] is the most convincing piece of production evidence that LLM-assisted diagnosis can actually lead to a reduction in time to diagnosis at a step in the recovery process. It's been live in Alibaba Cloud's global network for 10 months and has, on average, reduced time to root cause by 20.5% and by 55.2% for high-severity incidents. It achieved 95.5% top-1 accuracy on 357 real cases. Taken together these studies demonstrate that diagnostic quality is very much a function of dependency knowledge and telemetry and that is why the proposed framework is so focused on the enterprise data foundations.

4.2. Intelligent Service Delivery and Ticket Management

Mandal et al. [10] demonstrate how to implement enterprise-scale routing of email messages to the help desk using an ensemble classifier and a rule engine that can be programmed. It processed over 90k messages per month across three providers, achieved around 90% accuracy and coverage, and resulted in a massive reduction in manual effort. Service desk automation, combined with AI classification and operational rules, offers real benefits, with an eye to scalability and the ability to adapt to changing business needs. Outside of the tech industry, the only practitioner case study is from a tire company, Apollo Tyres (India) [27]. They deployed a pilot of generative AI for their IT service desk, which included multilingual user interactions, categorizing incidents from free text descriptions, predicting asset failures and a GenAI-powered FAQ. They were able to achieve a 30% reduction in manual interventions roughly two months after going live, and reduce the mean incident resolution time from 67.3 to 40.5 hours. It's a fairly straightforward before and after type of study over a very short timescale with no control group so the figures are very early indicators. What makes the case worth reading is what came before the pilot. A data governance foundation (data quality, access roles, audit trails, retention) was already in place. Frontline staff were trained first. Prompt workflows were designed together with the service desk, operations and audit teams, and the rollout began in the lower-risk areas. Accurate ticket routing is not in itself the same thing as successful resolution. A ticket can be correctly routed and still take a long time to resolve, requiring a lot of investigation, a lot of collaboration and a lot of expertise. The initial process that constitutes intelligent service delivery is the classification of tickets followed by the retrieval of knowledge, diagnostic support, resolution recommendations, escalation to human and learning from previous

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

incidents.

4.3. LLM-Assisted Root-Cause Analysis and Mitigation

Ahmed et al. [6] explored the performance of a collection of LLMs with respect to more than 40,000 production incidents across more than 1000 cloud services. Their comparison of zero-shot fine-tuned and multitask variants demonstrates that LLMs can be of assistance to on-call engineers in the areas of root cause analysis and mitigation planning. The approach is validated by a human evaluation with incident owners. RCACopilot as described by Chen, Y. et al. [18] operates by first automatically harvesting the diagnostic information via incident-specific handlers and then using an LLM to predict the root cause category and provide an explanation for that prediction. On a year's worth of Microsoft incidents it achieved a root cause category accuracy of 0.766. This matters for enterprise customers because the system's accuracy depends on collecting structured diagnostic evidence before invoking the LLM, not on the LLM alone. LLM assistance can save engineers a huge amount of time digging through historical records and putting together the various bits of evidence which are often all over the place. It can also make institutional knowledge much more accessible to the less experienced. Still, a generated recommendation is a hypothesis, not a diagnosis. The model sees only what it is given, and an incident usually turns on things that change by the hour: live telemetry, current dependencies, the configuration as it actually stands, and local knowledge nobody wrote down. Retrieval-augmented generation helps by tying answers to approved runbooks, knowledge articles, incident histories and postmortems, but only if those sources are current. A stale runbook, retrieved and presented with confidence, can do more harm than no runbook at all.

4.4. Agentic and Tool-Using Systems

Work published since 2024 has seen a move away from the single step LLM recommendation approach to agents that go out and gather data, invoke tools and reason over multiple steps. RCAgent [7] is an autonomous, tool-augmented agent running on an internally hosted model within Alibaba Cloud's real time compute platform for Apache Flink. It has beaten the ReAct baseline in both automated and human evaluation on root cause, solution, evidence and responsibility. The fact that it's using an internally hosted model rather than an external service is of course a governance decision to keep the incident data within the provider's boundary. The benchmark and testbed work is actually quite sobering. In ITBench [9], agents based on state-of-the-art models managed 11.4% of the site reliability engineering scenarios, 25.2% of the compliance/security scenarios, and 25.8% of the financial operations scenarios. Cornacchia et al. [28] found that LLM agents attempting end-to-end root cause analysis in a microservices testbed would fail via hallucinated reasoning paths and through inefficient use of context. All of that suggests that agentic AIOps is still very much in its infancy when it comes to being able to operate autonomously in the real world in a general purpose way. It's moving quite quickly in the narrow, deployed case but general purpose autonomous remediation is still not reliable enough to operate without human oversight.

4.5. Governance, Reliability, and Human Oversight

Hallucination and overconfidence are the two main risks with LLM-assisted diagnosis. Zhang et al. [11] address these by using LM-PACE which estimates the confidence of an LLM's root cause prediction by looking at how well it is supported by historical reference incidents. The confidence is calibrated so that engineers can distinguish between predictions that are well supported and those that aren't. A high confidence score doesn't necessarily mean the prediction is correct so confidence should be used in conjunction with evidence review and human approval. Overall, the governance structures adopted in the studies can be divided into two main categories. Some are explicitly evaluated: calibrated confidence [11], structured evidence collection prior to LLM inference [18] and consistency checks along the agent's reasoning path [7]. Some are just noted as part of their deployment without any individual assessment: use of an internally hosted model for data protection [7], data governance controls, audit at each stage of deployment, phased deployment and human escalation [27]. The standard controls (least privilege access, formal change approval, tested rollback) are not presented as having been evaluated in any of the primary studies. They are simply listed in Section 5 as being in line with industry standard practice and are appropriately annotated in Shows Table 4.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

Table 2 Characteristics of the primary studies included in the synthesis, in order of publication

Study	Setting	Focus and approach	Evidence type	Principal contribution	Limitation
Mandal et al. [10], 2019	IBM; 3 service providers	Ticket dispatch; ensemble classifier with rules	Production deployment	~90% accuracy and coverage; >90,000 messages/month	Routing only
Chen, Z. et al. [17], 2020 (IcM BRAIN)	Microsoft; 2-year study	Incident-management practice; AIOps platform	Empirical study	Key obstacles: incomplete dependencies, imprecise resource health	Single provider
Ghosh et al. [5], 2022	Microsoft; large cloud service	High-severity incidents; postmortem analysis	Production data	Incident causes, response delays, automation gaps	Single service; no AI evaluated
Ahmed et al. [6], 2023	Microsoft; >40,000 incidents, >1,000 services	LLM root-cause and mitigation suggestions (GPT-3/3.5)	Offline; incident-owner review	Useful root-cause and mitigation suggestions	Expert validation needed; no deployment
Jin et al. [19], 2023 (Oasis)	Microsoft; 18 cloud systems	Outage impact assessment; LLM summaries	Production data; outage-owner review	Outage scoping and summaries	Summaries in early adoption
Chen, Y. et al. [18], 2024 (RCACopilot)	Microsoft; 1 year of incidents	Diagnostic collection, then LLM root-cause prediction	Production incidents	Root-cause category accuracy up to 0.766	Categories only, not remediation
Zhang et al. [11], 2024 (LM-PACE)	Microsoft; cloud incidents	Retrieval-based confidence estimation for LLM RCA	Offline; production data	Calibrated confidence scores	Reduces, but does not remove, error
Wang, Z. et al. [7], 2024 (RCAgent)	Alibaba Cloud; Flink real-time compute platform	Tool-using LLM agent on an in-house model	Deployed; automated and human evaluation	Outperforms ReAct baseline	Single platform; relative gains only
Wang, C. et al. [8], 2025 (BiAn)	Alibaba Cloud; global network	LLM failure localization	Production; 10 months	Time to root cause cut 20.5% (55.2% for high-risk); 95.5% top-1 on 357 cases	Network domain; single provider
Jha et al. [9], 2025 (ITBench)	IBM; 102 scenarios	Agent benchmark: SRE, compliance, FinOps	Controlled benchmark	Solved 11.4% of SRE and ~25% of compliance and FinOps scenarios	Not production
Cornacchia et al. [28], 2025	KAUST; microservice testbed	LLM agents for end-to-end RCA	Testbed	Failure modes: hallucinated reasoning, poor use of context	Laboratory only
Apollo Tyres case [27], 2026	Manufacturer (India); IT service desk	GenAI for multilingual support, categorization, asset-failure prediction, FAQ	Case study; before/after metrics and surveys	Resolution time 67.3 → 40.5 h; ~30% fewer manual interventions	2-month window; no control group

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

Results in five main findings, and the research questions each answers are in brackets:

- AI can enhance some functions of ITSM (specifically in ticket classification, anomaly detection, diagnosis, outage assessment, and outage knowledge retrieval) (RQ1).
- These capabilities are most meaningful when viewed as a whole when integrated throughout the incident lifecycle, not as distinct models [RQ1, RQ3].
- LLMs and agents extend the scope of automation but introduce new reliability and governance issues. Autonomous agents continue to perform very poorly on end-to-end tasks [RQ4].
- Operational effectiveness is very much a function of data quality, dependency knowledge, system integration, organizational knowledge, workforce readiness and human expertise [RQ2].
- There is far more evidence of task-level performance than of sustained, enterprise-wide resilience – only two studies report time-based operational outcomes [RQ3].

4.6. Answers to the Research Questions

Table 3 shows the mapping of each research question to the themes, studies and findings, and the next few paragraphs describe the answer to each research question. As shown in Table 3 Mapping of research questions to themes, evidence, and answers.

Table 3 Mapping of research questions to themes, evidence, and answers

RQ	Themes and findings	Principal evidence	Answer in brief
RQ1	Themes 1–2; Findings 1–2	[6], [8], [10], [18], [19], [27]; workflow context [5], [17]	Partially. Task-level gains; two deployments show time savings; no enterprise-scale end-to-end evidence
RQ2	Theme 2; Finding 4	[5], [6], [17], [18], [27]	Context over model: telemetry, dependencies, integration, documented knowledge, skills
RQ3	Theme 3; Findings 2 and 5	[5], [8], [17], [19], [27]	Plausible but unproven. Faster response and retained learning; no longitudinal outcomes
RQ4	Theme 4; Finding 3	Evidence: [7], [9], [11], [18], [27], [28]; Proposed: Table 4	Evidence-based: confidence, structured evidence, in-house models, data governance, phased rollout, escalation. Proposed: least privilege, change control, rollback

RQ1: AI and incident resolution/service delivery. The evidence relates to RQ1 only in part. AI has been shown to improve certain aspects of the work but end-to-end improvements have only been demonstrated in a very small number of cases. Deployed ticket dispatch achieved circa 90% accuracy and coverage at scale [10]. LLM-based systems have been shown to generate useful root cause and mitigation recommendations on large volumes of production incidents [6], [18]. Outage impact assessment and summarization have been used to assist engineers during major incidents [19]. There are only two studies that report on outcomes in terms of time. One reports a 20.5% average reduction in time to root cause over a 10-month period in a production network [8]. The other reports a reduction in mean resolution time from 67.3 to 40.5 hours in an uncontrolled 2-month pilot [27]. The studies also indicate that a correct classification or recommendation will only result in a shorter recovery time if it is then fed into the next stage of the incident lifecycle [5], [17].

RQ2: Technical and organizational factors. Effectiveness is really more to do with the environment around the model than the model itself. The technical factors are telemetry quality, full service dependency information and structured diagnostic data collection [17], [18]. The organizational factors are consistent incident

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

documentation, up-to-date runbooks and knowledge articles, clear ownership and escalation paths, integration between ITSM and observability tools and people who can validate the AI outputs [5], [6]. The manufacturing case offers additional support that the improvements were a result of a foundation of data governance, training the workforce, co-design with all the relevant stakeholders, and a phased rollout [27]. The absence of this builds a false sense of security for the AI, as it will be re-classifying previous errors or making recommendations that are incorrect.

RQ3: Contribution to operational resilience. AI contributes to resilience in that it reduces the time between detection, diagnosis and mitigation, helps the engineers to understand the scope of an outage and turns past incidents and postmortems into reusable knowledge [5], [17], [19]. Two of the papers report on time-based outcomes [8], [27] but none report on a longitudinal effect on availability, downtime or incident recurrence at the enterprise level. If resilience is defined in the way it is in the resilience engineering literature (as the capacity of a system to anticipate, absorb, respond and recover from disruption [25]) then the evidence is largely around response and to a lesser extent learning. The contribution to resilience is therefore both plausible and mechanistically valid, just that it hasn't yet been demonstrated at the enterprise level.

RQ4: Governance mechanisms. The answer to RQ4 is to split the control into two categories. The reviewed studies support calibrated confidence with action thresholds [11]; gathering diagnostic evidence before calling an LLM [18]; consistency checks on the agent's own reasoning, and locally hosted models that keep incident data in house [7]; and data governance controls (audit at each rollout stage, phased deployment, human escalation) [27]. The benchmark and testbed results [9], [28] all support the idea of human authorization for high-risk actions. Least privilege access, formal change approval, simulation before execution, and tested rollback and execution tracing are not addressed in any primary study and are proposed in this paper. They are taken from standard security, service management and reliability practice [1], [2], [29], [30] and from established AI risk management frameworks [22] to [24]. Table 4 maps both types of control to the relevant activities by risk tier and provides the rationale for each.

5. Proposed Governance-Centered Framework

Figure 2 shows the resulting framework. Five layers are joined by solid arrows, which show the direction in which value flows. Governance sits underneath as a separate layer, and its dashed arrows mark where its controls apply.

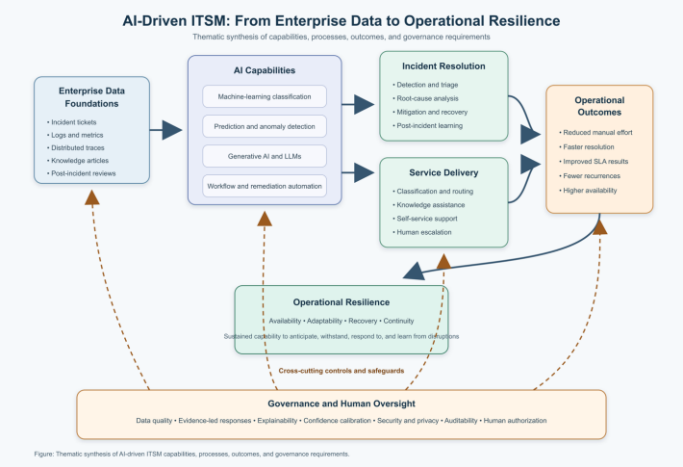

Figure 2 Governance-centric framework for AI-driven ITSM, linking the data foundations, the AI capabilities, the ITSM processes, the operational outcomes, and the operational resilience, with governance and human oversight as the cross-cutting moderator

5.1. Layers of the Framework

Enterprise data foundations (the left-hand side of Figure 2) are the inputs upon which every AI capability is predicated: incident tickets, logs and metrics, distributed traces, knowledge articles and post-mortems. The diagnostic studies [17], [18] and the manufacturing case [27] all show that it's the completeness and quality of this layer which ultimately caps everything that comes downstream.

The AI capabilities used in the various studies within the review are: machine learning for classification, prediction and anomaly detection; generative AI and large language models (LLMs); and workflow and remediation automation (including tool-using agents). ITSM processes are split into two streams. Incident resolution covers detection and triage, root cause analysis, mitigation and recovery and post-incident learning. Service delivery covers classification and routing, knowledge assistance, self-service and human escalation. That's the evidence in the studies [10], [27] on service desk performance and [6], [8], [18] on incident diagnosis which measure

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

different things and a gain in one stream doesn't necessarily equate to a gain in the other. Operational outcomes are what the business actually feels: less manual effort, faster resolution, better service levels, fewer repeat incidents and higher availability. Few of the reviewed studies get this far. Most stop at task metrics, and only two report time-based outcomes at all [8], [27]. Operational resilience is the ability to anticipate, absorb, respond to and learn from disruption [25]. The arrow from outcomes to resilience indicates that resilience is the outcome of sustained improvements in outcomes rather than the result of any one deployment. It is the layer for which the evidence is weakest.

5.2. Governance as a Moderator

The governance layer (bottom of Figure 2) is the collection of controls that are identified in Section 4.5 and in Table 4: data quality, evidence-led responses, explainability, confidence calibration, security and privacy, auditability and human authorization. The dashed arrows from this layer go to the data foundations, the AI capabilities, the ITSM processes and the outcomes. Moderation has a very specific meaning here. Governance in itself doesn't create value. It will be about the percentage of the enhancement that is transferred to the next and how well it is transferred. The data that is available to be input into the model will be based on data quality and privacy controls. The confidence thresholds and level of evidence that will be needed to take action on an AI output will dictate whether or not they are actually acted on. Whether or not the recommended action is implemented will depend on human authorization and change control. The deployment will be continued, adapted or discontinued based on the outcome monitoring. NIST AI RMF does not consider governance as an approval phase at the very end of the AI lifecycle but as a continuous process that is applied throughout the lifecycle [22] and ISO/IEC 42001 is an AI management system which spans an entire organization [24].

5.3. Testable Propositions

The framework suggests three propositions which can be tested in future empirical studies:

- P1. AI capability is necessary but not sufficient for better service outcomes; its effect also depends on data foundations and workflow integration.
- P2. Governance has a moderating effect on the interaction of capability and outcomes: when the activity is more risky, higher levels of governance will decrease the number of adverse automated actions taken, but will keep most of the time saved from AI assistance.
- P3. Resilience gains only occur when the improvements in outcomes are sustained and then fed back into the data foundations via post-incident learning.

6. Discussion, Limitations, and Future Research

6.1. Novelty and Contribution to Existing Research

The novelty of this work is in the way it positions AI within the enterprise rather than in the AI technique itself. As Table 1 illustrates, previous reviews [3], [4], [14] – [16] have classified AIOps by technique, by task and by data source. Our framework contributes three elements which the prior work only touches on. Firstly it connects the AI capabilities to the change of the ITSM processes and then to the service level and resilience outcomes, making the unit of evaluation restored service instead of model accuracy. Secondly, it understands governance as a moderator at each point of that chain instead of an independent last step and mapping governance controls with existing ITSM, security and AI governance standards. Thirdly it integrates LLM and agent specific risks (hallucination, confidence calibration, evidence grounding, unreliable autonomous action) into a single space as the more traditional AIOps tasks. The study also contributes to theory in that it positions AI-enabled ITSM as a sociotechnical system. The evidence supports two of the propositions in Section 5.3 directly: technical capability is a necessary but not sufficient condition for operational benefit and data foundations, workflow integration and human expertise all act as mediators in the relationship between AI capability and service outcomes. All three of the propositions could be tested in future empirical work. Finally, the review points to a deficiency in the evidence base itself. Of the 12 primary studies half were carried out in the cloud services of a single hyperscaler. The addition of studies from Alibaba Cloud, IBM, a manufacturer and an academic testbed does help to diversify the base but doesn't close the gap as none of them come from a regulated sector. The large enterprise environments in healthcare,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

financial services, and the public sector operate in a very different environment to that of the studies described. They are subject to multi-vendor escalation, have dependencies on legacy and mainframe systems, have formal change advisory and audit processes and have legal obligations around the handling of protected data. In this context automated remediation is constrained by contractual and regulatory obligations (e.g. DORA for EU financial institutions [20]) and the governance mechanisms identified in RQ4 become not just desirable but absolutely essential. The framework is designed with governance and the enterprise foundations in mind so that it can be applied in this context and validation in such an environment is a key priority for future work.

6.2. Discussion and Practical Implications

AI-based ITSM is more than just a technology deployment; it's a sociotechnical transformation. The studies reviewed do show genuine advances in classification, prediction, root cause analysis and recommendation but the value is entirely contingent on how well the capabilities are embedded in the processes and in human decision making. Adoption should move in steps, with risk setting the pace. Low-risk work such as ticket categorization, duplicate detection, knowledge retrieval and incident summaries can be automated early, provided someone reviews a sample of the output. Production configuration changes, service restarts, traffic redirection and automated remediation are another matter. They need authorization, confidence thresholds, testing or simulation, restricted permissions, progressive rollout and a tested rollback path. The benchmark evidence makes the point plainly: agents built on current models resolved only 11.4% of SRE scenarios in ITBench [9], which is not a record that justifies unsupervised production changes. An organization's incident history is one of its more valuable assets, though it is rarely treated that way. Postmortems, knowledge articles, monitoring data and old tickets are what an AI system learns from and what it retrieves when it suggests a fix, so their quality sets the ceiling. Thin incident descriptions, outdated runbooks, inconsistent categories and missing dependency information all pass straight into the model and come back out as confident-sounding advice. The Apollo Tyres case is instructive here: the data governance groundwork came before the reported gains, not after [27]. AI is a tool that is likely to alter the job of IT workers, rather than replace them. Routine activities can be automated – engineers can focus on verifying evidence, making riskier decisions and enhancing operational knowledge. Evidence from the manufacturing case indicates that intentional workforce preparation, such as training and addressing employee concerns over potential job loss is needed for this transition to succeed. AI-powered ITSM is not just an increase in model accuracy. If the system's output can't be connected to the right workflow, there's little reason to expect recovery time to fall, and an accurate root-cause recommendation is of little use if the responsible team lacks the permissions, resources or tested procedures to act on it. Operational resilience is the ability to identify, understand, coordinate, restore and learn from disruptions. Model risk can arise at any point during an incident, which is why the proposed framework gives enterprise foundations as much weight as AI capabilities and applies governance throughout.

6.2.1. Practical Implications

The sensible place to start is a specific, measurable operational problem, not a general mandate to "use AI". Ticket routing, alert deduplication, incident summarization, similar-incident retrieval and knowledge-article recommendation are good candidates: the risk is low, and the effect shows up quickly in routing accuracy, reassignment rate, MTTA, MTTR, analyst effort and SLA attainment. Mandal et al.'s dispatch system, which handled more than 90,000 messages a month [10], shows that this kind of work already runs at enterprise scale. Organizations should have a set of baseline measurements before implementing any changes so they can then evaluate the difference between existing and improved, if they make any changes. The AI-generated recommendations should include the supporting evidence for the recommendation, the level of confidence in the recommendation and the data on which the recommendation is based. Organisations should also be clear as to what actions the AI can take (automated), what actions need to be approved (human) and what actions are prohibited (blocked). Table 4 presents a risk tiered adoption model which implements these principles. The automation modes are based on the levels of automation model of Parasuraman et al. [13] and the outcome measures of high-risk changes (change

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

failure rate, mean time to recover, etc.) are adopted from existing software delivery research [31].

The applications, type of incidents, operational language and infrastructure dependencies change continuously and continuous monitoring is required after deployment. Monitor the performance of the models to determine if they are degrading in performance, changing in data distribution, displaying new errors, or having unintended impacts on operating. The models and the workflows themselves should be continually improved based on feedback from the service desk and the operations engineers.

Table 4 Risk-tiered adoption model for AI-driven ITSM. "Evidence" marks controls supported by the reviewed primary studies; "Proposed" marks controls drawn from standards that the reviewed studies have not evaluated.

Risk tier	Example ITSM activities	Automation mode	Minimum controls	Evaluation metrics	Basis of controls
Low	Ticket categorization, duplicate detection, knowledge retrieval, incident summarization	Automated, with sampled human review	Periodic accuracy review; feedback capture; privacy filtering of incident text	Routing accuracy, reassignment rate, analyst effort	Evidence [10], [27]; Proposed [23]
Medium	Prioritization, resolver-group assignment, root-cause suggestions, mitigation recommendations	AI recommends; human decides	Evidence citation; calibrated confidence with action thresholds; audit trail of recommendations	Time to acknowledge, time to diagnose, recommendation acceptance rate	Evidence [11], [18], [27]
High	Configuration changes, service restarts, traffic redirection, automated remediation	Human authorization; progressive rollout	Change-control approval; least-privilege permissions; simulation or testing; tested rollback; execution tracing	Mean time to recover, change failure rate, incident recurrence, service-level compliance	Proposed [1], [2], [29], [30]; Indirect evidence [9], [28]
Restricted	Actions on protected data or safety-critical systems outside approved scope	Not automated	Policy enforcement; access denial; escalation to accountable owner	Policy violations (target: zero)	Proposed [20], [21], [24]

6.3. Limitations

The study has several limitations. It is based on a synthesis of 12 primary studies and 5 secondary sources which were selected purposively rather than through a full systematic screening process (hit counts and screening numbers were not recorded so the exact selection cannot be reproduced). Relevant work may have been missed and the studies do not cover all technologies, all industries or all operating environments. The conclusions should therefore be viewed as an evidence-informed synthesis rather than a complete assessment of the AIOps field. Secondly, most significant studies are based on proprietary data from the big tech firms. This data is of great value in terms of providing insight into real-world production environments, but it's not something that can be independently verified, and the results may not be transferable to firms with less complex infrastructure, less mature operations, less incident history, or different regulatory regimes. Thirdly the studies reviewed use very different datasets, different tasks, different experimental set-ups and different evaluation criteria. Ticket assignment accuracy can't be compared with root cause recommendation quality or with reductions in recovery time. This precludes statistical meta-analysis and requires a thematic synthesis instead. Fourthly, there's publication bias,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

which will tend to favor the successful AI implementations. The unsuccessful projects, the inaccurate models, the abandoned deployments and the general resistance to AI within organizations are not going to be reported in the peer-reviewed literature. The published evidence therefore may present a far more positive picture of the impact of AI on ITSM than the reality for the majority of organizations that have adopted it. Fifthly, AIOps and generative AI are moving at a fast pace. The search window was closed in October 2026, and in the meantime new models may emerge, new agentic systems and evaluation methods may appear, and new governance requirements may follow. The proposed framework is based on published evidence, but hasn't yet been verified in an independent case study, so empirical evidence is needed before determining causality. Sixthly, most evidence is from one organizational context, as six of the twelve key studies were carried out in the cloud services of one hyperscaler, and no studies are from a regulated sector. The only study from outside the tech sector is a very brief uncontrolled pilot. The findings will not necessarily be generalizable to organizations with multi-vendor arrangements or operating in a more regulated environment. Seventhly, The author is employed by IBM and two of the principal studies in the paper ([9], [10]) were authored by IBM researchers. The inclusion criteria were the same as for all the other sources though the reader should be aware of the affiliation. Some of the characterizations, particularly of the secondary sources in Table 1, are based on the abstract and stated scope rather than on full text analysis.

6.4. Future Research

The most pressing need is for longitudinal studies inside enterprises, measured before and after AI is introduced and compared against a control group. Model accuracy should still be reported, but next to the numbers operations teams actually track: resolution time, recurrence, downtime, SLA compliance, user satisfaction and engineering effort. Without that, the field will keep reporting task-level gains that no one can tie to service outcomes. Cross-sector research is also needed. The financial services, health, manufacturing, telecoms, government, and retail sectors all operate in very different risk and regulatory environments, and comparative case studies could help establish which AI capabilities and governance approaches are transferable across sectors and which are necessarily sector-specific. Work in the EU financial services sector would be a particularly interesting application of the framework as DORA has very specific requirements around ICT risk management and incident reporting [20]. Human-AI collaboration during high severity incidents is something that absolutely warrants far more attention. There's a need for research into whether engineers understand the AI confidence information, when they will reject an AI suggestion, how the level of automation impacts situational awareness [12] and whether less experienced staff become overly reliant on AI explanations. Standard AIOps benchmarks and evaluation criteria (ITBench [9] is in this direction) and anonymised public incident datasets are all required to improve reproducibility. The evaluation needs to consider false recommendations, hallucinations, response latency, model drift, security, explainability and the operational impact of errors. Agentic AIOps systems which pull the telemetry, select the tools, run their own diagnostics and then go on to suggest or perform the remediation should be evaluated in a controlled environment prior to being deployed into production. The safety research in this area should cover bounded permissions, policy enforcement, execution tracing, human approval, progressive rollout and auto-rollback and should report on the failure modes observed [28]. Finally, a governance framework proposed here must be validated by conducting interviews with experts, surveys, case studies or a design science evaluation. Input from service desk, SREs, IT managers, security personnel, risk officers and business stakeholders would be helpful to fine tune the individual components and inter-dependencies. The propositions in Section 5.3 provide the basis for hypothesis testing.

Conclusion

This paper looks at the potential for AI to transform incident management, service delivery and operational resilience in large enterprises. The literature review shows that AIOps can be used to process massive volumes of operational data to assist with incident detection, failure prediction, ticket classification, root cause analysis, mitigation recommendations and automated remediation. ML is particularly good at classification and pattern recognition, and the arrival of large language models has brought a whole raft of new capabilities to the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

table in terms of understanding unstructured incident data, retrieving operational knowledge and producing appropriate recommendations. Agentic systems go further still but are still not reliable on end-to-end tasks. The review concludes on four points. With respect to RQ1 AI does improve certain ITSM functions (ticket routing, triage, diagnosis, outage assessment, knowledge retrieval) in production. End to end improvement has only been demonstrated in two of the deployments and not at enterprise scale. With respect to RQ2 the benefits are contingent upon reliable telemetry, complete dependency information, good incident records, up to date knowledge bases, integration into the operational workflow, workforce readiness and human expertise. For RQ3 AI contributes to operational resilience by reducing the time between detection and mitigation and by preserving the operational learning, though there's still very limited enterprise level evidence on availability and incident recurrence. For RQ4 the risks of hallucination, incorrect diagnosis, privacy exposure, security vulnerabilities, lack of explainability and over-reliance on automation all require evidence of grounding, calibrated confidence, data governance and human escalation (all of which the reviewed work does address) and also least privilege access, change control and tested rollback (which this paper proposes on the basis of existing standards). The paper's contribution is fourfold: a comparative positioning vis-a-vis other AIOps reviews; a lifecycle view of the empirical evidence and a risk tiered adoption model that organizations can use straight away; a governance based framework that brings together the enterprise data foundations, the AI capabilities, the ITSM process improvements, the service outcomes and the resilience aspects; and a research agenda which is outcome focused and cross industry in its scope. The evidence base is very much rooted in hyperscale cloud operations so the next step is to validate the framework in the more regulated, multi-vendor enterprise environment. AI shouldn't be assessed on its ability to automate individual ITSM tasks alone. The real value to the enterprise is in the degree to which it enables services to be restored faster, delivered more consistently and operated with greater resilience without in any way compromising accountability or control. AI ITSM works best when intelligent automation and human judgment are seen as co-existing and complementary. The organizations that are going to get the most out of AIOps are those that combine the technical innovation with good data management, good process integration and sound risk-based governance.

Conflict of interest. The author is employed by IBM as a Senior Consultant (Application Development). IBM develops and sells AIOps and IT-automation products. Two of the included primary studies [9], [10] were authored by IBM researchers, the author was not involved in either study. The author declares no other competing interests.

Disclaimer. The views expressed in this paper are the author's own and do not necessarily represent those of IBM

References

[1]. AXELOS. (2019). ITIL Foundation: ITIL 4 edition. TSO.
[2]. International Organization for Standardization. (2018). Information technology—Service management—Part 1: Service management system requirements (ISO/IEC 20000-1:2018).
[3]. Remil, Y., Bendimerad, A., Mathonat, R., & Kaytoue, M. (2024). AIOps solutions for incident management: Technical guidelines and a comprehensive literature review. arXiv.
[4]. Cheng, Q., Sahoo, D., Saha, A., Yang, W., Liu, C., Woo, G., Singh, M., Savarese, S., & Hoi, S. C. H. (2023). AI for IT operations (AIOps) on cloud platforms: Reviews, opportunities and challenges. arXiv.
[5]. Ghosh, S., Shetty, M., Bansal, C., & Nath, S. (2022). How to fight production incidents? An empirical study on a large-scale cloud service. In Proceedings of the 13th ACM Symposium on Cloud Computing (SoCC '22).
[6]. Ahmed, T., Ghosh, S., Bansal, C., Zimmermann, T., Zhang, X., & Rajmohan, S. (2023). Recommending root-cause and mitigation steps for cloud incidents using large language models. In Proceedings of the 45th International Conference on Software Engineering (ICSE) (pp. 1737–1749).
[7]. Wang, Z., et al. (2024). RCAgent: Cloud root cause analysis by autonomous agents with tool-augmented large language models. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24).

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

[8]. Wang, C., et al. (2025). Towards LLM-based failure localization in production-scale networks. In Proceedings of the ACM SIGCOMM 2025 Conference (SIGCOMM '25).

[9]. Jha, S., et al. (2025). ITBench: Evaluating AI agents across diverse real-world IT automation tasks. In Proceedings of the 42nd International Conference on Machine Learning (ICML), 267, 27134–27197.

[10]. Mandal, A., Malhotra, N., Agarwal, S., Ray, A., & Sridhara, G. (2019). Automated dispatch of helpdesk email tickets: Pushing the limits with AI. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 9381–9388.

[11]. Zhang, S., Zhang, X., Bansal, C., Las-Casas, P., Fonseca, R., & Rajmohan, S. (2024). LM-PACE: Confidence estimation by large language models for effective root cause analysis of cloud incidents. In Companion Proceedings of the 32nd ACM International Conference on Foundations of Software Engineering (FSE Companion '24).

[12]. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779.

[13]. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30(3), 286–297.

[14]. Dang, Y., Lin, Q., & Huang, P. (2019). AIOps: Real-world challenges and research innovations. In Proceedings of the 41st International Conference on Software Engineering: Companion Proceedings (pp. 4–5).

[15]. Notaro, P., Cardoso, J., & Gerndt, M. (2021). A survey of AIOps methods for failure management. ACM Transactions on Intelligent Systems and Technology, 12(6), Article 81.

[16]. Zhang, L., et al. (2025). A survey of AIOps in the era of large language models. ACM Computing Surveys.

[17]. Chen, Z., et al. (2020). Towards intelligent incident management: Why we need it and how we make it. In Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20).

[18]. Chen, Y., et al. (2024). Automatic root cause analysis via large language models for cloud incidents. In Proceedings of the 19th European Conference on Computer Systems (EuroSys '24) (pp. 674–688).

[19]. [19]. Jin, P., et al. (2023). Assess and summarize: Improve outage understanding with large language models. In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '23).

[20]. European Union. (2022). Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (Digital Operational Resilience Act). Official Journal of the European Union, L 333, 1–79.

[21]. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[22]. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).

[23]. National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1).

[24]. International Organization for Standardization. (2023). Information technology—Artificial intelligence—Management system (ISO/IEC 42001:2023).

[25]. Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.). (2006). Resilience engineering: Concepts and precepts. Ashgate.

[26]. Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (Technical Report EBSE-2007-01). Keele University and Durham University.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5883 - 5901
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0760

[27]. Biswas, B. (2026). Generative AI implementation in enterprises: Lessons from a case study of enhanced IT service management. Information Systems Journal.

[28]. Cornacchia, A., Alabdulaal, I., Saghier, I., Mirdad, A., Fayoumi, O., & Canini, M. (2025). Between promise and pain: The reality of automating failure analysis in microservices with LLMs. In Proceedings of the 16th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys '25) (pp. 155–167).

[29]. Joint Task Force. (2020). Security and privacy controls for information systems and organizations (NIST Special Publication 800-53, Rev. 5). National Institute of Standards and Technology.

[30]. Beyer, B., Jones, C., Petoff, J., & Murphy, N. R. (Eds.). (2016). Site reliability engineering: How Google runs production systems. O'Reilly Media.

[31]. Forsgren, N., Humble, J., & Kim, G. (2018). Accelerate: The science of lean software and DevOps. IT Revolution.