Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 10| 2025

Research Article **Received:** 20-09-2025| **Accepted:** 15-10-2025 | **Published:** 31-10-2025

Scaling GenAI Agent Assistants: A Production Framework for Enterprise Customer Support Systems

Amaan Javed

Independent Researcher, USA

Abstract: The article suggests a systematic approach to the implementation of generative AI agent assistants into enterprise customer support solutions to fill the long-standing gap between prototype solutions and production-ready solutions. Going beyond one-off technical aspects, the framework inculcates foundational human experience obtained through golden dataset techniques, addresses the inter-system integration challenges, and identifies clear metrics of success. The phased implementation plan is described, which includes pre-production validation, scalable architecture design, and controlled launch processes. The technical architecture is concerned with integration details, performance optimisation, and data governance, which are important when deploying to an enterprise. A holistic evaluative approach balances early warning indicators with lagging business impact assessment indicators, enhanced with operational health indicators. Empirical evidence demonstrates the presence of measurable results and qualitative insights, as well as several unforeseen observations, which highlight the fact that the generative AI solutions are evolving in response to complex business situations.

Keywords: Enterprise Genai Implementation, Customer Support Transformation, Human-AI Collaboration, Measurement Framework, Production Architecture.

INTRODUCTION

The Production Gap Challenge

The generative AI (GenAI) ecosystem is defined by the large gap between the flashy displays and the real-world enterprise application. Existing empirical research on AI use in business processes suggests that it is estimated that about 65-70 per cent of organisations cannot go beyond initial experimentation and that the more large-scale the implementation, the greater the challenges (Merhi, M. I. 2023). This gap in production is particularly pronounced in the customer support operation, but complex webs of old systems, compliance requirements, and dependencies of operations create complexities not be replicated by demonstration environments. Enterprise environments require solutions that can integrate effectively with the existing infrastructure and maintain consistent performance even with variable load requirements and conditions, conditions which are radically different from controlled showcase environments.

Large-scale solutions driven by large-language-model solutions have organisational, economic, and governance aspects that are not confined to technical complexities. Studies of the slow uptake of technology have shown that GenAI deployments at production scale require a high level of cross-departmental coordination, coordination systems that are not present during pilot projects (Haefner, N . 2023). Organisations often do not consider the requirements of infrastructure to deploy enterprise functions, especially in the area of computational power and monitoring systems. The accelerated development of the foundation models causes decision paralysis in most businesses because of the fear of becoming obsolete due to the technology and the urgency to make an immediate move.

Contextualizing Enterprise Support Requirements

Enterprise customer support environments are unique with unique constraints as compared to consumer applications. Such environments generally deal with complicated product ecosystems that are typified by massive configuration variation, requiring support systems that can be used to travel through multi-dimensional settings that may be written across technical specifications and account history. Research on the digital transformation of customer service has demonstrated that, as an enterprise support interaction, there are usually many stakeholders with conflicting interests that present complexities on an unprecedented scale in B2C situations (Merhi, M. I. 2023).

The economic impact of the quality of support goes beyond operational conditions into affects the underlying business relations. According to the research on B2B customer experience, support quality is one of the key factors that define the possibility of retaining and expanding an account, particularly in the context of subscription-based models. Enterprise support interactions are made at

***Corresponding Author:** Amaan Javed

DOI- https://doi.org/10.5281/zenodo.17492689

critical points, such as implementation issues, operation-related issues, or strategic decisions; therefore, they have a disproportionate impact on relationship patterns. This economic fact reinvents support as a cost centre to a strategic function that has a direct impact on retention of revenue (Haefner, N . 2023).

RESEARCH OBJECTIVES

A holistic model of operational GenAI implementation fills the critical gaps in research and practice. The framework development provides a systematic approach that is flexible to discrete operational settings but still offers the rigor to implement on a large enterprise scale. This framework focuses on end-to-end requirements, including data preparation, integration architecture, and governance mechanisms, unlike those that focus on model performance in controlled settings (Merhi, M. I. 2023).

The methodical approach to validation, architecture, and controlled implementation is the key to successful enterprise deployments. Methodologies of validation recommend a quantitative assessment plan to measure responses in more than one dimension. Architectural considerations are concerned with particular requirements, including compatibility with legacy systems and long-lasting load performance. The measurement framework integrates both operational measures and business results, developing a holistic perspective of the effect of implementation on the results in the long-term and allowing organisations to identify leading indicators that predict future performance (Haefner, N . 2023).

FOUNDATION: HUMAN EXCELLENCE AS BASELINE

Golden Dataset Methodology

To build useful systems of AI generation, it is necessary to establish a preliminary stable platform that is based on exemplary human interactions instead of average performance rates. The empirical researches on the digitization of expertise in service settings highlight the importance of choosing interactions that exhibit better results on a variety of measurement variables such as resolution efficacy, communicative clarity, emotional intelligence, and efficiency measures (Ledro, C. *et al.*, 2025). Such a multi-dimensional strategy will reduce the possibility of continuing with interactions that are efficient but lack emotional resonance or empathic but technically shallow.

The methods of annotation represent one of the determinants of system effectiveness. The studies that question the methodologies of knowledge transfer have found that the exhaustive annotation schemes include the overt content and the subtly covert patterns of decision-making, which is the foundation of the outstanding performance. These systems use rubric-based assessment models that have clear-cut standards on all the dimensions, thus promoting consistency and allowing subtle aspects of performance to be expressed (Batz, A. *et al.*, 2025). Organizations that have systematic systems of annotation governance, including guidelines, calibration sessions, and performance monitoring, build significantly more consistent underpinnings of model building.

Comparison analysis explains special trends in human and automated responses. The professional human agent shows advanced contextual adaptation, including gradual disclosure, implicit knowledge exploitation, and managing the flow of conversation in a balance that is efficient and comfortable to the customer. Studies of human-AI cooperation have shown that high-performing agents' capabilities spontaneously apply adaptive strategies that need to be clearly formalized in automated systems (Ledro, C. *et al.*, 2025).

Multi-System Integration Challenges

Enterprise support environments cut across a number of operational systems, thus occasioning considerable integration issues. Research on digital transformation points out that the boundaries of systems tend to mirror organizational structures of the past, not based on customer-focused design principles, but each system only maintains some customer context that is needed to provide full support (Ledro, C. *et al.*, 2025). This fragmentation presents a problem to AI systems that are generative and require context to create the right response.

Taxonomic harmonization is one of the most problematic areas, and it requires the resolution of conflicting classification systems. Studies of enterprise knowledge management show that taxonomic inconsistencies are frequently reflective of sound operational variations and not just standardization failure (Batz, A. *et al.*, 2025). Proper strategies add the intermediate layers of translation that keep the system-specific taxonomies intact and allow the unified analysis via carefully mapped connections.

Publisher: SARC Publisher

The concept of temporal alignment adds more complexity to the use of historical support data. Empirical observations show that support strategies for the same problem go through a tremendous change over the years as new products are developed, policies redefined, or new best practices emerge (Ledro, C. *et al.*, 2025). Their robust strategies should thus include the means of determining temporal relevance, significant changes in approaches, and ensuring the training examples are weighted accordingly.

Table 1: Foundation - Human Excellence as Baseline (Ledro, C. *et al.*, 2025; Batz, A. *et al.*, 2025)

Component	Key Elements	Purpose
Golden Dataset Methodology	Exemplary interactions selection, Annotation schemas, Rubric-based assessment	Establishes a performance benchmark based on the best human agents
Multi-System Integration	Taxonomic harmonization, Translation layers, Temporal alignment	Addresses fragmentation across operational systems
Metric Definition	Hero metrics, Experience metrics, Correlation analysis	Balances operational efficiency with customer experience

Metric Definition and Success Parameters

The development of the relevant metrics requires a certain distinction between the indicators of operational efficiency and the quality of experience. Research concerning customer experience management indicates that the vast majority of organizations tend to overvalue easily measurable operation-related factors at the cost of more complex dimensions of experience (Ledro, C. *et al.*, 2025). Competent models make a difference between hero measures of headline performance and experience measures of quality dimensions across the customer journey.

The statistical correlation analysis gives invaluable input to the system prioritization. A survey of customer experience analytics reveals complicated links between operational measures, experience measures, and business deliverables, such as retention, growth, and recommendation (Batz, A. *et al.*, 2025). Such interdependencies are often non-linear, have threshold effects, inflection points, and interaction effects.

Controlled experimentation requires the establishment of a baseline, which must take into consideration inherent variation. Literature on experimental design illustrates support environments to have special challenges due to the variability of customer needs, relative complexity of the issue, and large numbers of confounding variables (Ledro, C. *et al.*, 2025).

THE THREE-PHASE IMPLEMENTATION FRAMEWORK

Phase 1: Pre-Production Validation

Effective GenAI implementations for enterprises need to be systematically pre-produced and must go beyond the traditional testing procedures. Research on organizational use of artificial intelligence affirms that organized validation systems are essential preconditions that facilitate performance thresholds, besides identifying possible limitations before implementation. Multidimensional assessment models built upon comprehensive evaluation matrices reflect the complex nature of enterprise support environments, combining aspects like factual accuracy, contextual appropriateness, policy compliance, and linguistic quality (Rezazadeh, A. *et al.*, 2025).

Iterative prompt engineering techniques are an essential part of validation and allow systematic refinement by using systematic experimentation. Empirical research on the deployment of AI in enterprises emphasizes that structured prompt engineering is a separate field of study that requires specialized skills in the technical areas and organization standards of communication. Proper approaches introduce version-controlled experimentation, which is carefully documented, and the particular alteration, together with the equivalent performance effects, are recorded and, as such, generate evidence-based engineering choices as opposed to a collection of intuitive enhancements (Veeravalli, P. K. 2025).

Categorization taxonomies on failure help to perform classification of system limitations according to their severity, as well as the method of their resolution. Strong taxonomies distinguish between critical failures that are characterized as material risks, significant failures that affect experiential quality, and minor problems that require refinement but do not affect core functionality (Rezazadeh, A. *et al.*, 2025). Groundedness thresholds also require a statistical validation, which provides critical performance

Publisher: SARC Publisher

assurances, especially in the regulated setting, and powerful strategies utilize stratified sampling strategies amongst unequal content types (Veeravalli, P. K. 2025).

Phase 2: Architecture for Scale

Scalable architecture is an important success factor of implementation into the enterprise that has to support a high volume of transactions. The Buildversusbuy decision model is one of the basic considerations that considers the commercial platforms solutions versus in-house development strategies in terms of performance features, integration functionality, flexibility in customization, and financial aspects (Rezazadeh, A. *et al.*, 2025).

The middleware orchestration patterns are the key elements of dealing with large transaction volumes, which involve request handling, processing optimization strategies, and operation protection (Veeravalli, P. K. 2025). The performance engineering techniques enable organizations to achieve the needed throughput, as well as maintain the acceptable response times, in a staged manner, starting with the creation of a baseline, moving into systematic optimization, and finally concluding with full load testing (Rezazadeh, A. *et al.*, 2025).

Data governance frameworks will make sure that it is implemented and meets the security and compliance criteria without adversely affecting operational efficiency. Sophisticated schemes are provided on data security, preservation of privacy, and compliance with regulations, between high controls and operational flexibility (Veeravalli, P. K. 2025).

Phase 3: Controlled Launch Strategy

Planned launch plans provide organized ways of moving systems to production without causing a lot of disturbance. Sequential implementation procedures introduce gradual exposure patterns that progressively escalate usage using validated performance (Rezazadeh, A. *et al.*, 2025). The designs of A/B experiments confirm performance in the live conditions, maintaining the statistical rigor, and applying controlled comparison experiments between the existing strategies and GenAI functionality (Veeravalli, P. K. 2025).

Setting the integration mechanisms of feedback builds the input of agents and customers into the refinement of the systems and employs multimodal feedback mechanisms that involve structured agent input, customer experience measurement, and operational analytics (Rezazadeh, A. *et al.*, 2025). Economic modeling provides analytical models used in the assessment of expansion decisions based on proven performance, which include direct operational effects, quality improvement, and customer experience value addition (Veeravalli, P. K. 2025).

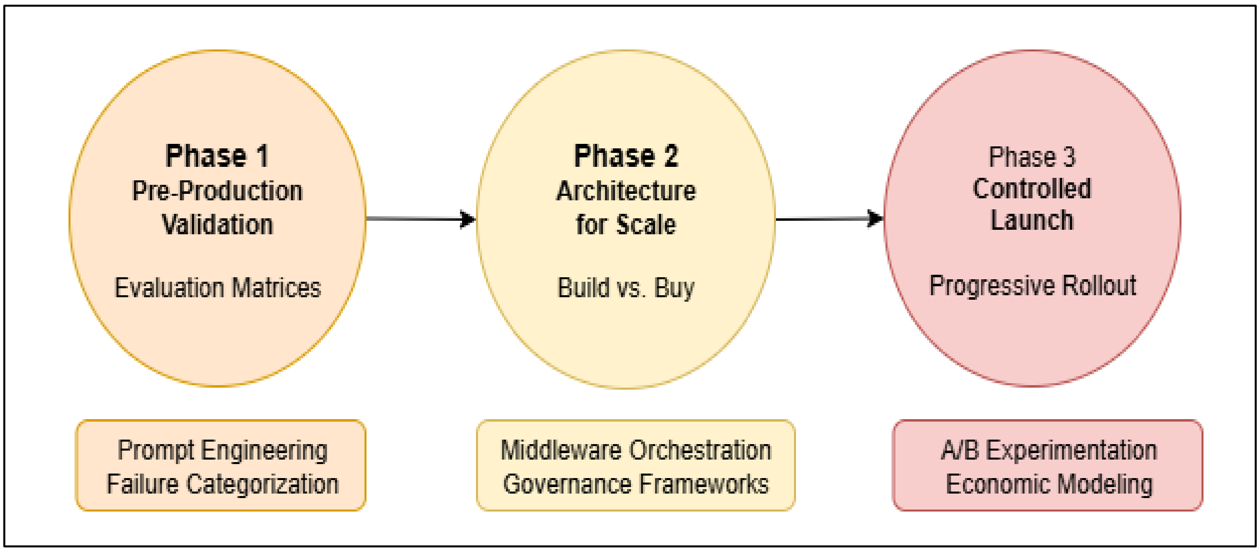

Fig 1: Three-Phase Implementation Framework (Rezazadeh, A. *et al.*, 2025; Veeravalli, P. K. 2025)

TECHNICAL ARCHITECTURE DESIGN

Integration Architecture Considerations

The architectural approach is the basic determinant of integration of generative artificial intelligence capabilities in enterprise systems, which defines functional capabilities and performance characteristics. Empirical studies of enterprise-scale generative AI systems indicate that integration choices map out system boundaries that are hard to change after implementation. A

Publisher: SARC Publisher

comparative study of the middleware and the native integration strategies has shown that there are unique benefits realized according to the organizational context, with no single integration strategy being superior to the rest. Native integrations help simplify the implementation and make it less complex, whereas middleware architectures enable more control over transaction management, error handling, and enforcement of governance, benefits that are more valued as the implementation scale grows (Ettinger, A. 2025).

Another important design decision that has important implications is the decision between pull and push architectural models. Pull architectures locate the enterprise systems in the position of interaction controllers, thus compromising the ability to govern at a higher level, with the possible disadvantage of real-time responsiveness. On the other hand, push architectures introduce more event-based models of interaction that are capable of making them more responsive, but limit the ability to control when to transact with the enterprise system. Such an architectural choice must thus not only take into account the technical aspects but also organizational aspects, such as governance needs, and preferred operational control (Rashid, A. B., & Kausik, M. A. K. 2024).

In the implementation of workflow optimization, there is the aim of balancing responsiveness and throughput in an enterprise. Synchronous processing can provide instant consistency and easier error management, but asynchronous processing can offer higher throughput and efficiency of resource utilisation. Good architectures deploy workflow selection that is purpose-specific, i.e., matching processing patterns to operational needs, as opposed to using generic strategies (Ettinger, A. 2025). Templates Prompt cache and processing result techniques produce substantial performance improvements at lower operation costs in high-volume implementations (Rashid, A. B., & Kausik, M. A. K. 2024).

Performance Engineering

Transaction processing optimisation is an important engineering emphasis of enterprise GenAI implementations that can support high volumes. The quality throughput required is achieved by ensuring systematic optimisation in connection management, request batching, and parallel processing capabilities (Ettinger, A. 2025). Latency optimisation algorithms maintain responsive user experiences in real-time systems by targeting network transmission, computational efficiency, and response handling improvements that cumulatively enhance responsiveness as perceived (Rashid, A. B., & Kausik, M. A. K. 2024).

Opposite, resource utilisation optimisation can be used to run the computational economics of GenAI implementations by performing systematic efficiency improvements in compute allocation, model selection, and workload distribution (Ettinger, A. 2025). Error handling and resilience patterns maintain operational stability in mission-critical environments by reducing the effects of temporary faults, dependency failures, and capacity constraints, and thus sustaining vital functionality in sub-optimal operating conditions (Rashid, A. B., & Kausik, M. A. K. 2024).

Data Governance Implementation

It becomes a governance requirement that there are security controls for the sensitive information. The effective security implementation ensures the overall data life cycle by controlling access, transmissions, data security in storage, and data security in operation (Ettinger, A. 2025). Privacy compliance systems govern the data processing operations that have different regulatory needs, use data minimisation, purpose restriction, storage policies, and subject-rights control (Rashid, A. B., & Kausik, M. A. K. 2024).

The implementation of retention policy offers formalized systems of managing data life cycle from acquisition to disposal, and audit facilities can show the effectiveness of control and operation within controlled settings through the documentation of policies, activities are logged, compliance management, and the management of the evidence (Ettinger, A. 2025).

Publisher: SARC Publisher

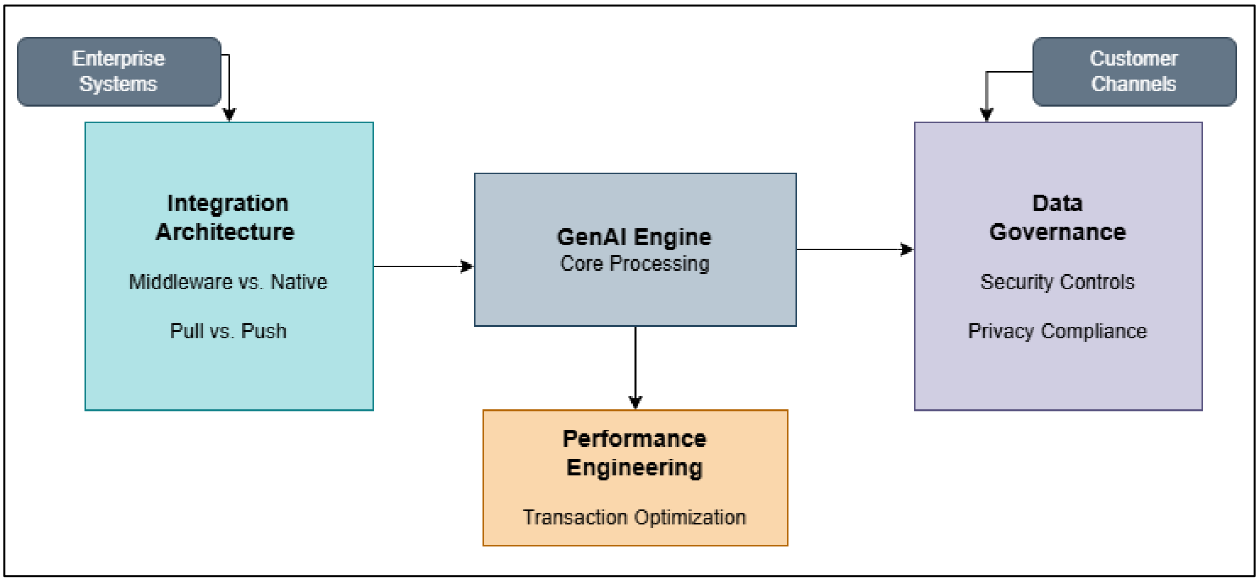

Fig 2: Technical Architecture Design (Ettinger, A. 2025; Rashid, A. B., & Kausik, M. A. K. 2024)

MEASUREMENT FRAMEWORK

Leading Indicator Methodology

The development of effective leading indicator practices helps organisations predict the performance of a given system in the future before the business is affected downstream. The body of research that examines measurement frameworks of intelligent systems highlights the importance of leading indicators as the key early warning systems that can transform management from a reactive remediation approach to proactive optimisation. Good frameworks utilize multidimensional monitoring, which is based on technical, operational, and experiential areas as opposed to using individual metrics that can ignore crucial indicators (Tominc, P. *et al.*, 2023).

The measurement of agent adoption provides preliminary indicators of system adoption and value conception by the primary users. Evidence based on human-AI collaboration shows that implementation trends are usually followed by other measures of performance, where behavioural changes in the use of AI are the early warning signs to an impending issue before showing up in other performance or business measures. In-depth measurement goes further than the number of uses to look at the intensity of interaction, the variety of feature use, and the stability of engagement (Rabhi, F. *et al.*, 2025).

Quality assessment of summary generation gives insight into a back-end capability that has many downstream implications. The study of knowledge-intensive processes emphasizes the fact that the quality of information synthesis at the outset triggers a set of consequences in the course of further actions. The entire framework considers such dimensions as factual accuracy, contextual relevance, structural organisation, and communicative clarity-that is, a set of elements that cumulatively produce meaningful value (Tominc, P. *et al.*, 2023).

Early signal detection systems detect possible problems at an early stage of appearance, which provides an opportunity to intervene in advance. It has been shown that even small changes in patterns often give way to apparent degradation. Efficient structures have in place the daily monitoring of stability measures, content quality measures, and patterns of interaction, as it is realised that issues do not tend to become dramatic crashes but quite often manifest as some slight deviation (Rabhi, F. *et al.*, 2025).

Lagging Indicator Analysis

Strong lagging indicator analysis reaffirms the final business impact, that is, technical gains are turned into organisational value. Studies indicate that detailed frameworks allow organisations to identify the creation of value at various levels that might not be visible when analysis is done based on operational measures (Tominc, P. *et al.*, 2023).

Measurement of escalation reduction provides information regarding efficiency and improvement of the experience of operations. Wholesome strategies examine complexity allocation, theme classification, and chronology to indicate whether enhancements show uniformity across the types of issues or vary over time (Rabhi, F. *et al.*, 2025).

The validation of customer satisfaction provides essential data concerning the improvement of the

Publisher: SARC Publisher

experience. Good methodologies use multidimensional frameworks that include direct feedback, behavioural signals, comparative evaluation, and relationship-strength measures to establish which of the experience dimensions motivates change (Tominc, P. *et al.*, 2023).

The evaluation of resolution time gives evidence on the efficiency improvement of the economic and the experience quality. Complexity correlation, resolution completeness, and interaction efficiency will be analysed in complex ways to ascertain whether quality will be preserved at the instances of speed improvements (Rabhi, F. *et al.*, 2025).

Operational Health Monitoring

Full operational health monitoring provides insight into the performance of the system, allowing proactive operations. Good frameworks use ongoing evaluation on a variety of dimensions as opposed to sampling occasionally (Tominc, P. *et al.*, 2023).

Performance monitoring API in real time captures the response-time distribution, throughput capacity, error rate, and utilisation of resources. In categorisation of errors, distinction is made according to the severity of impact, persistence, and patterns of failure, which allows suitable response measures in line with operational importance (Rabhi, F. *et al.*, 2025).

Anomaly detection detects a possible problem before it becomes overtly failed, and applies statistical modelling, pattern recognition, and correlation analysis to detect an outlier between metrics, whereas individual values are nominal (Tominc, P. *et al.*, 2023).

The categorisation of feedback transforms the input of users into a set of priorities to be acted upon and classifies the input into types of issues, impact measurement, frequency measurement, and implementation considerations to build an improvement roadmap based on the real-life user experience (Rabhi, F. *et al.*, 2025).

Table 2: Measurement Framework & Empirical Results (Tominc, P. *et al.*, 2023; Rabhi, F. *et al.*, 2025)

Category	Key Elements	Observed Outcomes
Leading Indicators	Agent adoption, Quality assessment, Signal detection	Early warning system for performance issues
Lagging Indicators	Escalation reduction, Satisfaction metrics, Resolution time	Business impact validation
Operational Health	API monitoring, Error categorization, Anomaly detection	Enables proactive management
Qualitative Insights	Agent feedback, Usage patterns, Cultural factors	Reveals adaptation and adoption factors
Unanticipated Discoveries	Complex account handling, Resource management, Workflow transformation	Highlights the evolutionary nature of implementation

EMPIRICAL RESULTS AND OBSERVATIONS

Quantitative Outcomes

The implementation of generative AI agent assistants in enterprise customer-support systems has produced significant, empirically quantifiable improvements in the most fundamental performance indicators. Extensive deployment studies show that there is a steady trend of increase in operations after well-planned implementation procedures. The agents will witness efficiency gains in a variety of aspects, i.e., less research time, acceleration of case processing, and simplification of resolution methodology. Such advantages are usually conforming to nonlinear adoption curves, in which smaller gains in the first parts of the acclimatization are followed by a speeding up of the improvement as the patterns of use increase. Mid-career practitioners are also more likely to have stronger growth effects as compared to both beginners and more highly tenured practitioners, which is consistent with the idea that generative capabilities have a specific value to agents with sufficient domain knowledge to move around systems, but before they acquire exhaustive personal expertise (Challapally, A. *et al.*, 2025). The indicators of customer satisfaction following the implementation indicate positive trends across a variety of measurement frameworks. Medium complexity interactions and moderate information synthesis interactions always exhibit the greatest improvements, with both extremely simple transactions and extremely complex situations responding with more limited improvements. This allocation is rationally consistent with the features of capabilities, where straightforward interactions provide little scope for enhancing capabilities, whereas complicated

Publisher: SARC Publisher

scenarios can exceed the technological standards of the present day (Collins, C. *et al.*, 2021). The decrease in the escalation rates is a highly beneficial result, a combination of the gains in operational efficiency with the gains in the experiential quality. The variability of the types of escalations is found in categorical analysis wherein procedural escalations indicate the greatest improvement, followed by technical relationship-driven escalations- an order that is similar to the prevailing technological competencies, where knowledge-based situations make use of the information retrieval and synthesis strengths (Challapally, A. *et al.,* 2025).

Qualitative Insights

The use of agent-feedback methodologies has given important qualitative insights into the impact of the implementation, which is critical compared to the quantitative measures. Thematic analyses always show the following patterns: the confidence-acceleration theme (increased readiness to solve rather complicated problems), the cognitive-offloading theme (decreased mental exhaustion as a result of delegating information retrieval activities), and the consistency-enhancement theme (increased standardization of responses resulting in predictable customer experiences) (Challapally, A. *et al.*, 2025). Deliberate model usage patterns have also not come as expected, as the adaptations go beyond the expected usage patterns. Regular patterns are preparation pattern (examining the similar cases of the past before interacting), verification pattern (authenticating recommended solutions before offering), and the translation pattern (transforming technical data to a client-appropriate language) (Collins, C. *et al.*, 2021).

The cultural aspects show a strong impact on the implementation outcomes of the technical consideration. Psychological safety is associated with a faster rate of adoption and an increased level of feature use. Learning orientation is associated with more agent-suggested enhancements, whereas a collaborative attitude is associated with higher implementation satisfaction (Collins, C. *et al.*, 2021).

Unanticipated Discoveries

Complex account cases have also become unexpected difficulties that require specific consideration. Merged case situations create challenges to context management, and hierarchical account structure presents challenges to policy application as it should. Such cases usually demand better context management, better entity recognition, and relationship modeling (Challapally, A. *et al.*, 2025). Mechanisms of resource management have become critical operational controls to maximize implementation economics. Graduated access models, which limit usage according to proven business value, bring significant cost savings without affecting the performance because restricting low-value over-utilization without reducing resources to essential functions (Collins, C. *et al.*, 2021). Patterns of workflow-transformation. The implementation patterns used naturally initiate the evolution of processes beyond mere automation of tasks. Patterns are preparation rebalancing (moving to proactive context building), verification restructuring (including solution validation before delivery), and documentation integration (including summary generation during the process of interaction) (Collins, C. *et al.*, 2021).

CONCLUSION

The proposed system of scaling generative AI agent assistants in the enterprise customer support setting gives a coherent framework that is comprehensive, as it deals with the multidimensional dilemmas of production implementation. With human excellence as a foundational pillar to system development, stringent validation processes, and scalable architectures that have proper governance controls, organisations can ensure that the gap between prototypical systems and working systems is bridged to meet the needs of large-scale enterprises. The measurement methodology enables the creation of a continuous optimisation process by balancing the leading and lagging measures, and the process promotes sustainable improvement cycles that are adjusted to the changing business demands. Although implementation issues are still ongoing, especially in the context of dealing with more complex accounts and resource optimisation, the framework still provides systematic recommendations on how to overcome these difficulties and still focus on business performance, as opposed to technical capacity. As the generative AI technologies keep on improving, this base provides a working discipline to support enterprise-level applications, which generate a positive business value and handle the risks thereof.

REFERENCES

1. Merhi, M. I. "An evaluation of the critical success factors impacting artificial intelligence

Publisher: SARC Publisher

implementation." *International Journal of Information Management* 69 (2023): 102545.

2. Haefner, N., Parida, V., Gassmann, O., & Wincent, J. "Implementing and scaling artificial intelligence: A review, framework, and research agenda." *Technological Forecasting and Social Change* 197 (2023): 122878.
3. Ledro, C., Nosella, A., Vinelli, A., Dalla Pozza, I., & Souverain, T. "Artificial intelligence in customer relationship management: A systematic framework for a successful integration." *Journal of Business Research* 199 (2025): 115531.
4. Batz, A., D'Croz-Barón, D. F., Vega Pérez, C. J., & Ojeda-Sanchez, C. A. "Integrating machine learning into business and management in the age of artificial intelligence." *Humanities and Social Sciences Communications* 12.1 (2025): 1-20.
5. Rezazadeh, A., Kohns, M., Bohnsack, R., António, N., & Rita, P. "Generative AI for growth hacking: How startups use generative AI in their growth strategies." *Journal of Business Research* 192 (2025): 115320.
6. Veeravalli, P. K. "Cloud-Native AI Solutions: Transforming enterprise application development." *World Journal of Advanced Research and Reviews,* (2025).
7. Ettinger, A. "Enterprise Architecture as a Dynamic Capability for Scalable and Sustainable Generative AI adoption: Bridging Innovation and Governance in Large Organisations." *arXiv preprint arXiv:2505.06326* (2025).
8. Rashid, A. B., & Kausik, M. A. K. "AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications." *Hybrid Advances* 7 (2024): 100277.
9. Tominc, P., Oreški, D., & Rožman, M. "Artificial intelligence and agility-based model for successful project implementation and company competitiveness." *Information* 14.6 (2023): 337.
10. Rabhi, F., Beheshti, A., & Gill, A. "Business transformation through AI-enabled technologies." *Frontiers in Artificial Intelligence* 8 (2025): 1577540.
11. Challapally, A. *et al.,* "The Genai Divide State Of AI In Business 2025." *MIT,* (2025).
12. Collins, C. *et al.,* "Artificial intelligence in information systems research: A systematic literature review and research agenda." *ScienceDirect,* (2021).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Javed, A. "Scaling GenAI Agent Assistants: A Production Framework for Enterprise Customer Support Systems." *Sarcouncil Journal of Engineering and Computer Sciences* 4.10 (2025): pp 322-330.

Publisher: SARC Publisher