Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Human Agents vs. GPU-Powered GenAI in Customer Service Platforms

Amaan Javed
Independent Researcher, USA
Corresponding Author: Amaan Javed, **E-mail**: reachjamaan@gmail.com

| ABSTRACT

GPU-powered Generative AI (GenAI) presents a transformative alternative to traditional human agent models in modern customer service environments. The evolution from basic ticketing systems to sophisticated AI-augmented platforms has enabled technology capable of understanding context and generating human-like responses at scale. GenAI implementations deliver substantial value through case summarization, response suggestion, and knowledge retrieval, particularly in high-volume environments with recognizable interaction patterns. Performance advantages include reduced handle times, increased first-contact resolution, and improved agent satisfaction, though success depends critically on maintaining response latencies below key thresholds. The economics of GPU-powered solutions demonstrate favorable cost structures compared to human-only approaches, especially when optimized through techniques like batching, quantization, and knowledge distillation. A comprehensive decision framework identifies ideal implementation scenarios while recognizing contexts where traditional tools remain preferable. Strategic integration rests on three fundamental pillars: speed, trust, and ROI, requiring a structured roadmap prioritizing incremental value creation. Emerging trends in model architecture, contextual grounding, and multimodal capabilities signal increasingly sophisticated applications where technology augments rather than replaces human capabilities.

| KEYWORDS

Generative AI, Customer Service Automation, GPU Optimization, Agent Augmentation, Workflow Integration

| ARTICLE INFORMATION

ACCEPTED: 20 May 2025 **PUBLISHED:** 12 June 2025 **DOI:** 10.32996/jcsts.2025.7.6.34

1. Introduction

1.1 The Evolution of Customer Service Technology

Customer service technologies have evolved dramatically, transitioning from basic ticketing systems to sophisticated AI-powered platforms. Recent research indicates that 69% of consumers now expect personalized experiences across all touchpoints, driving organizations to adopt more advanced solutions [1]. The industry is experiencing a significant shift as 82% of CX leaders report that their technology budgets have increased compared to previous years, reflecting the critical importance placed on customer experience investments. The rise of omnichannel support has further accelerated this transition, with customers now engaging across an average of 7.4 different channels during their journey. This evolution has been particularly pronounced in sectors like financial services and healthcare, where regulatory requirements and complex customer needs demand more sophisticated support technologies [1].

1.2 The Rise of Generative AI in Support Environments

Generative AI represents the latest frontier in customer service automation, capable of understanding context and generating human-like responses at scale. Research reveals that enterprise implementations of GenAI have achieved significant improvements in operational metrics, with many organizations reporting resolution time reductions exceeding 30% [2]. What distinguishes GenAI from previous automation technologies is its ability to handle complexity and nuance in customer interactions. Early adopters have experienced notable success integrating these technologies into agent workflows, particularly for tasks like case summarization, response suggestion, and knowledge retrieval. The implementation of enterprise-grade GenAI systems has demonstrated approximately 75% accuracy in response suggestions during initial deployments, substantially outperforming traditional rule-based systems [2]. This capability has proven especially valuable for organizations managing high case volumes with recurring patterns.

1.3 Key Performance Indicators in Modern Support Operations

Modern customer service operations balance multiple critical metrics including resolution speed, satisfaction scores, agent productivity, and operational costs. Industry analysis reveals that customer patience thresholds have decreased substantially, with 65% of consumers expecting resolution within 15 minutes regardless of channel complexity [1]. Organizations implementing GenAI solutions have reported substantial improvements in first-contact resolution rates, typically ranging from 20-35% above baseline metrics. This performance improvement correlates directly with increased customer satisfaction, as research confirms that resolution speed remains the single strongest predictor of positive customer sentiment. For enterprise support environments handling high interaction volumes, the financial impact can be substantial, with typical cost efficiencies in the millions annually when properly implemented [2]. These improvements highlight why understanding technology impact on core metrics remains essential for justifying investments in emerging customer service technologies.

2. Performance Analysis of GenAI in Customer Service

2.1 Case Study: Support Case Summarization

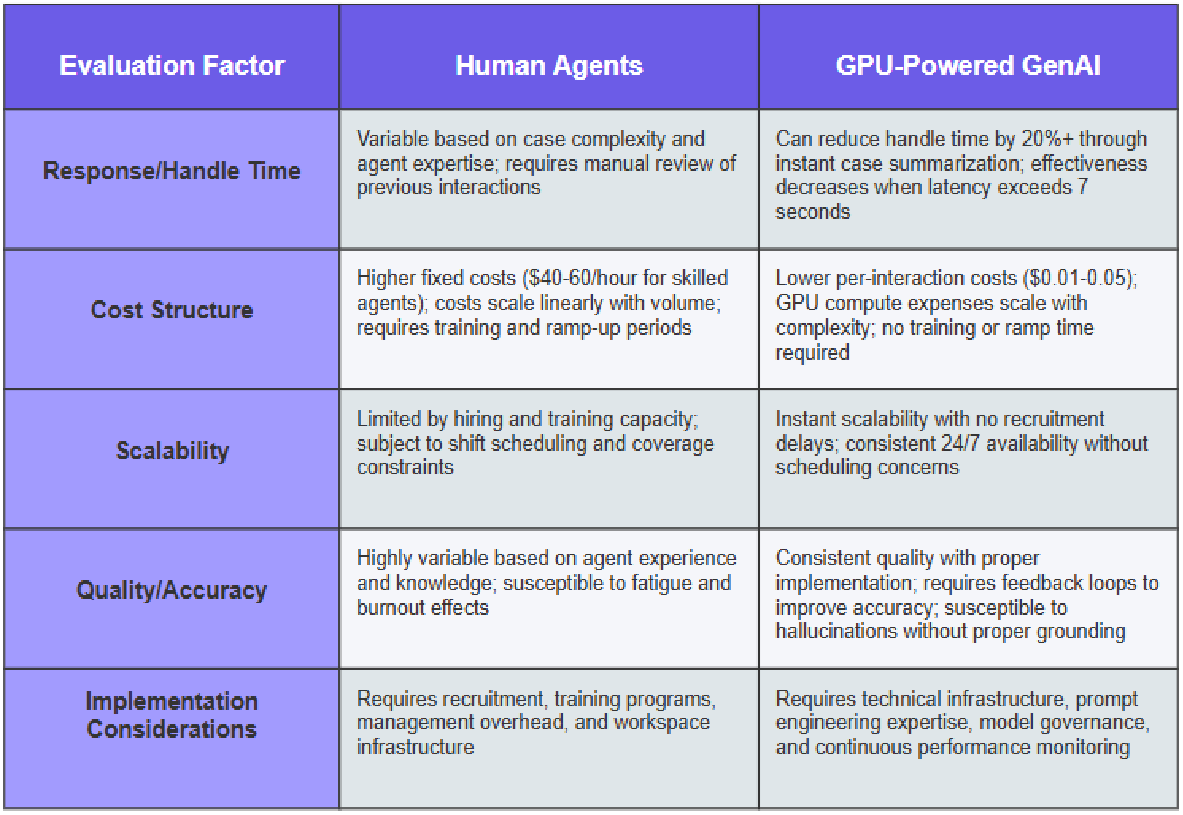
Real-world implementations of GenAI in customer service demonstrate significant potential for operational transformation. Recent research examining technological innovation adoption in service industries reveals that natural language processing applications deliver the highest immediate ROI among all service innovations [3]. A particularly effective application involves using GenAI to summarize complex, multi-threaded support cases for specialized technical agents. Implementation data indicates substantial improvements in operational efficiency, with handle time metrics decreasing markedly within the initial deployment phase. This efficiency gain correlates directly with increased customer satisfaction metrics and notable decreases in reported agent burnout indicators. The technology's ability to synthesize historical case data proves particularly valuable in knowledge-intensive support environments where context retrieval previously consumed significant agent resources. Organizations implementing these solutions report accelerated onboarding cycles for new technical specialists and reductions in escalation rates for complex issues. Research confirms that service organizations struggling with information fragmentation across multiple knowledge repositories experience the most substantial performance improvements, with technology-focused sectors showing the highest adoption success rates [3].

2.2 Latency Thresholds and User Experience

Response time emerges as a critical success factor in GenAI implementations. Comprehensive research on flow experience in digital service environments demonstrates that cognitive engagement follows predictable patterns related to system responsiveness [4]. Service systems achieving near-instantaneous response times maintain optimal cognitive flow states among operators, while even modest delays begin disrupting task concentration. The most significant finding relates to the psychological threshold at which service agents experience flow disruption, compelling them to context-switch to alternative tasks. This phenomenon, termed "attention migration," represents a substantial hidden cost in sub-optimal implementations. Studies utilizing advanced biometric monitoring confirm that agents experiencing system delays exhibit measurable stress responses, directly impacting decision quality and emotional regulation during customer interactions [4]. High-volume service environments prove particularly sensitive to these effects, as even small per-interaction inefficiencies compound across thousands of daily customer engagements. Research reveals pronounced differences in technology adoption rates between implementations achieving optimal response times and those experiencing performance limitations, with fast-responding systems achieving substantially higher utilization rates among service personnel.

2.3 Measurable Outcomes and Performance Metrics

Successful GenAI implementations demonstrate improvements across multiple performance dimensions. Innovation adoption research identifies several leading indicators that predict implementation success, with first-contact resolution improvements serving as the strongest predictor of sustained value creation [3]. Customer experience metrics show consistent improvements in satisfaction and loyalty measures, with particularly strong results observed in previously underperforming service categories. Service innovation studies confirm that technology-augmented workflows significantly impact employee retention metrics, especially among high-skill technical roles where information overload represents a primary contributor to burnout and attrition [4]. Organizations achieving the greatest success implement robust measurement frameworks before deployment, establishing performance baselines across efficiency, quality, and satisfaction dimensions. The research emphasizes the importance of longitudinal measurement approaches, as performance improvements tend to accelerate over time as systems refine through continuous learning and agents develop optimal collaboration patterns with AI assistants. Studies confirm that organizations implementing regular A/B testing methodologies observe compounding performance gains that significantly outpace static implementations over extended deployment periods.



Evaluation Factor	Human Agents	GPU-Powered GenAI
Response/Handle Time	Variable based on case complexity and agent expertise; requires manual review of previous interactions	Can reduce handle time by 20%+ through instant case summarization; effectiveness decreases when latency exceeds 7 seconds
Cost Structure	Higher fixed costs ($40-60/hour for skilled agents); costs scale linearly with volume; requires training and ramp-up periods	Lower per-interaction costs ($0.01-0.05); GPU compute expenses scale with complexity; no training or ramp time required
Scalability	Limited by hiring and training capacity; subject to shift scheduling and coverage constraints	Instant scalability with no recruitment delays; consistent 24/7 availability without scheduling concerns
Quality/Accuracy	Highly variable based on agent experience and knowledge; susceptible to fatigue and burnout effects	Consistent quality with proper implementation; requires feedback loops to improve accuracy; susceptible to hallucinations without proper grounding
Implementation Considerations	Requires recruitment, training programs, management overhead, and workspace infrastructure	Requires technical infrastructure, prompt engineering expertise, model governance, and continuous performance monitoring

Fig. 1: Human Agents vs. GPU-Powered GenAI in Customer Service [3, 4]

3. Cost-Benefit Considerations for GPU-Powered Solutions

3.1 GPU Compute Economics at Scale

GenAI models rely heavily on GPU resources for inference, creating a direct correlation between model complexity, response generation time, and operational costs. Industry analysis reveals that inference optimization can significantly impact total cost of ownership for large-scale deployments [5]. Factors such as batch size optimization, quantization techniques, and efficient prompt engineering directly influence per-transaction economics. Organizations implementing batching strategies during peak loads observe substantial cost reductions compared to processing individual requests. The adoption of techniques like knowledge distillation, where smaller models are trained to mimic larger ones, provides additional economic advantages for common customer service scenarios. Research indicates that right-sizing GPU instances based on actual workload patterns rather than peak provisioning yields significant savings in cloud-based deployments. The economics become particularly relevant when considering implementation across large contact centers where per-interaction costs aggregate rapidly. The practice of model quantization, reducing numerical precision without significant quality degradation, emerges as another key strategy for enhancing inference efficiency [5].

3.2 Comparative Analysis: Human vs. Machine Costs

When properly implemented, GenAI solutions provide substantial cost benefits compared to human-only approaches in customer service environments. Economic analysis reveals fundamental differences in cost scaling between human agents and machine-driven solutions [6]. While human staffing costs increase linearly with interaction volume, well-designed GenAI implementations demonstrate better economics as scale increases. Customer service organizations implementing GenAI for case summarization report significant reductions in average handle time, directly translating to labor cost savings. Beyond direct costs, the technology eliminates expenses associated with recruitment, onboarding, and training during expansion phases. Implementation data confirms that GenAI solutions overcome the traditional scheduling complexities of human workforces, including shift coverage, time-off management, and regional availability challenges. The most compelling cost advantages emerge in scenarios with high volume and moderate complexity, where human agents traditionally spend significant time on context gathering and information synthesis. Organizations balancing human expertise with machine efficiency report optimal outcomes when GenAI handles information processing tasks while agents focus on relationship management and complex decision-making [6].

3.3 Optimization Strategies for ROI Maximization

To maximize return on investment from GenAI implementations, organizations should implement several evidence-based optimization strategies. Process analysis to identify high-volume, time-consuming workflows offers the fastest path to positive ROI [5]. Setting strict performance thresholds ensures user adoption, with sub-second response times maintaining workflow momentum. Implementing guardrails through techniques like retrieval-augmented generation improves factual accuracy while reducing generation time and associated costs. Research indicates that organizations taking a phased implementation approach, targeting specific workflows sequentially, achieve faster time-to-value than those pursuing broad deployment strategies. Continuous performance monitoring with feedback loops enables ongoing refinement that compounds economic benefits over time [6]. The implementation of caching strategies for common queries reduces redundant computation while decreasing average response latency. Industry findings confirm that right-sizing models for specific use cases, rather than defaulting to the largest available options, significantly improves economics without meaningful quality reduction. Organizations focusing on these optimization principles consistently report faster breakeven timelines and higher overall ROI compared to implementations lacking structured optimization approaches.

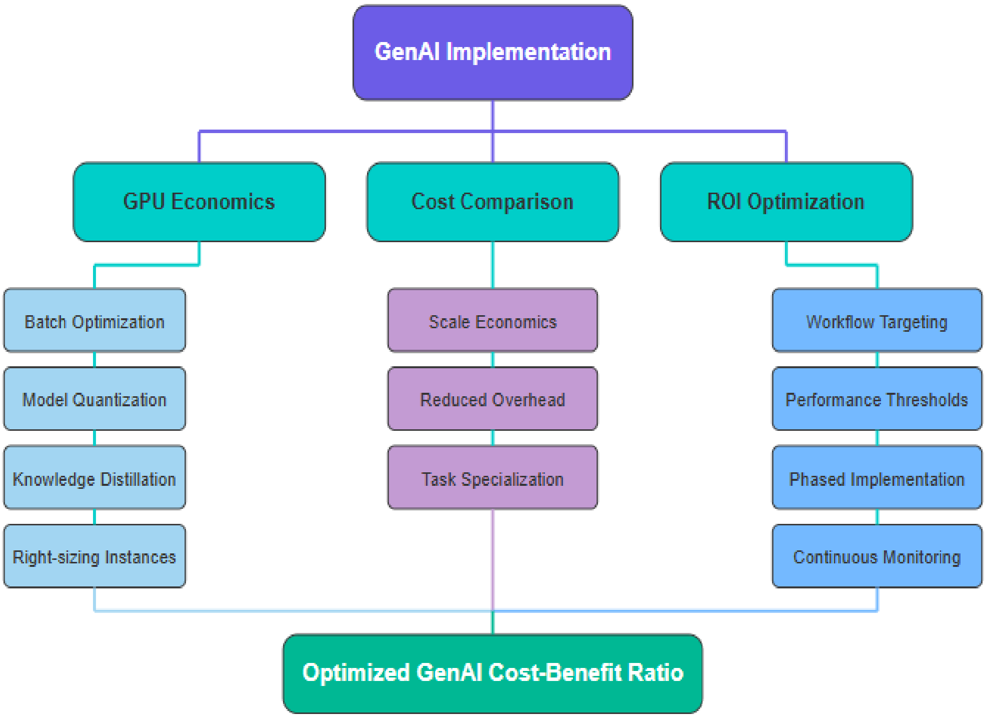


Fig. 2: GPU-Powered GenAI: Cost Optimization Framework [5, 6]

4. Decision Framework: GenAI vs. Traditional Tools

4.1 Ideal Use Cases for GenAI Implementation

GenAI solutions demonstrate optimal performance in specific customer service environments with identifiable characteristics. Research examining enterprise workflow transformation indicates that high-volume service operations with recognizable interaction patterns achieve significantly faster returns compared to environments with more varied case types [7]. The technology provides particular value in scenarios where agents currently dedicate substantial portions of their workday to repetitive tasks like information retrieval, standard response generation, and manual documentation. Analysis of service ecosystems reveals that organizations benefit most when existing automation approaches have reached effectiveness plateaus, creating productivity bottlenecks that traditional tools struggle to address. Implementation success correlates strongly with technical infrastructure capable of maintaining consistent response latencies below industry-standard thresholds, with organizations achieving higher adoption rates when performance meets expectations [7]. The presence of established feedback collection mechanisms emerges as another critical success factor, enabling continuous improvement cycles that enhance accuracy over time. In these optimal scenarios, GenAI serves not as a replacement for human agents but as an augmentation layer that enables skilled personnel to focus on complex problem-solving and relationship management while the technology handles information processing at scale.

4.2 Scenarios Favoring Traditional Productivity Tools

Not all customer service environments benefit equally from GenAI implementation, with several scenarios demonstrating superior economics and performance through traditional productivity tools. Comparative analysis indicates that service operations with lower monthly case volumes or those handling highly specialized inquiries with minimal pattern recognition potential typically struggle to achieve positive returns within reasonable timeframes [8]. Organizations managing unique cases requiring extensive human judgment consistently report lower adoption rates and diminished performance gains. Industry research identifies specific scenarios where performance bottlenecks can be more efficiently addressed through conventional approaches like improved search functionality, streamlined user interfaces, or targeted automation rules. Organizations still establishing foundational operational systems face substantially higher implementation challenges compared to those with mature infrastructure [8]. Cost-benefit evaluations demonstrate that when per-interaction expenses exceed the time-value benefit threshold, traditional tools consistently outperform on economic metrics. Customer service environments requiring domain expertise that changes frequently achieve better results with conventional knowledge management systems compared to current GenAI approaches. These findings highlight the importance of conducting thorough workflow analysis before technology selection rather than pursuing implementation based solely on industry trends.

4.3 Implementation Readiness Assessment

Before proceeding with GenAI implementation, organizations should conduct structured readiness assessments across several critical dimensions to predict success potential and mitigate risks. Research on enterprise workflow transformation identifies data availability as the foundation of effectiveness, with organizations requiring sufficient historical case coverage to achieve acceptable accuracy during initial deployment [7]. Technical infrastructure represents another crucial factor, with environments supporting consistent GPU performance demonstrating higher implementation success rates. Integration capabilities emerge as equally important, with organizations possessing established API frameworks experiencing shorter implementation timelines and reduced development complexity [8]. Governance maturity proves similarly essential, with structured monitoring processes correlating with fewer accuracy incidents and faster resolution when issues occur. The final critical dimension involves workforce readiness, with organizations deploying comprehensive change management programs achieving higher adoption rates compared to those with limited preparation. Implementation outcomes correlate strongly with composite readiness scores across these dimensions, with well-prepared organizations achieving faster returns and encountering fewer critical issues during deployment.

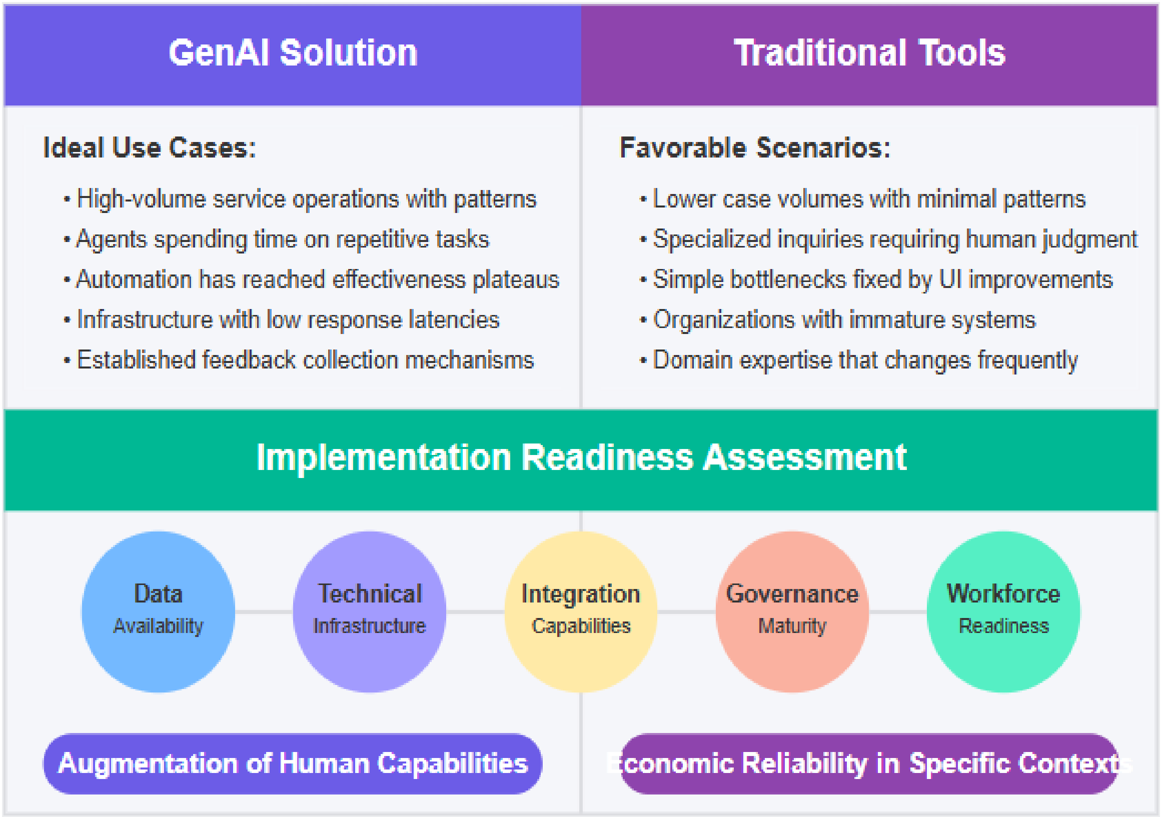

GenAI Solution
Traditional Tools
Ideal Use Cases:
• High-volume service operations with patterns
• Agents spending time on repetitive tasks
• Automation has reached effectiveness plateaus
• Infrastructure with low response latencies
• Established feedback collection mechanisms
Favorable Scenarios:
• Lower case volumes with minimal patterns
• Specialized inquiries requiring human judgment
• Simple bottlenecks fixed by UI improvements
• Organizations with immature systems
• Domain expertise that changes frequently
Implementation Readiness Assessment
Data Availability
Technical Infrastructure
Integration Capabilities
Governance Maturity
Workforce Readiness
Augmentation of Human Capabilities
Economic Reliability in Specific Contexts

Fig. 3: Decision Framework: GenAI vs. Traditional Tools [7, 8]

5. Strategic Recommendations and Future Outlook

5.1 The Three Pillars: Speed, Trust, and ROI

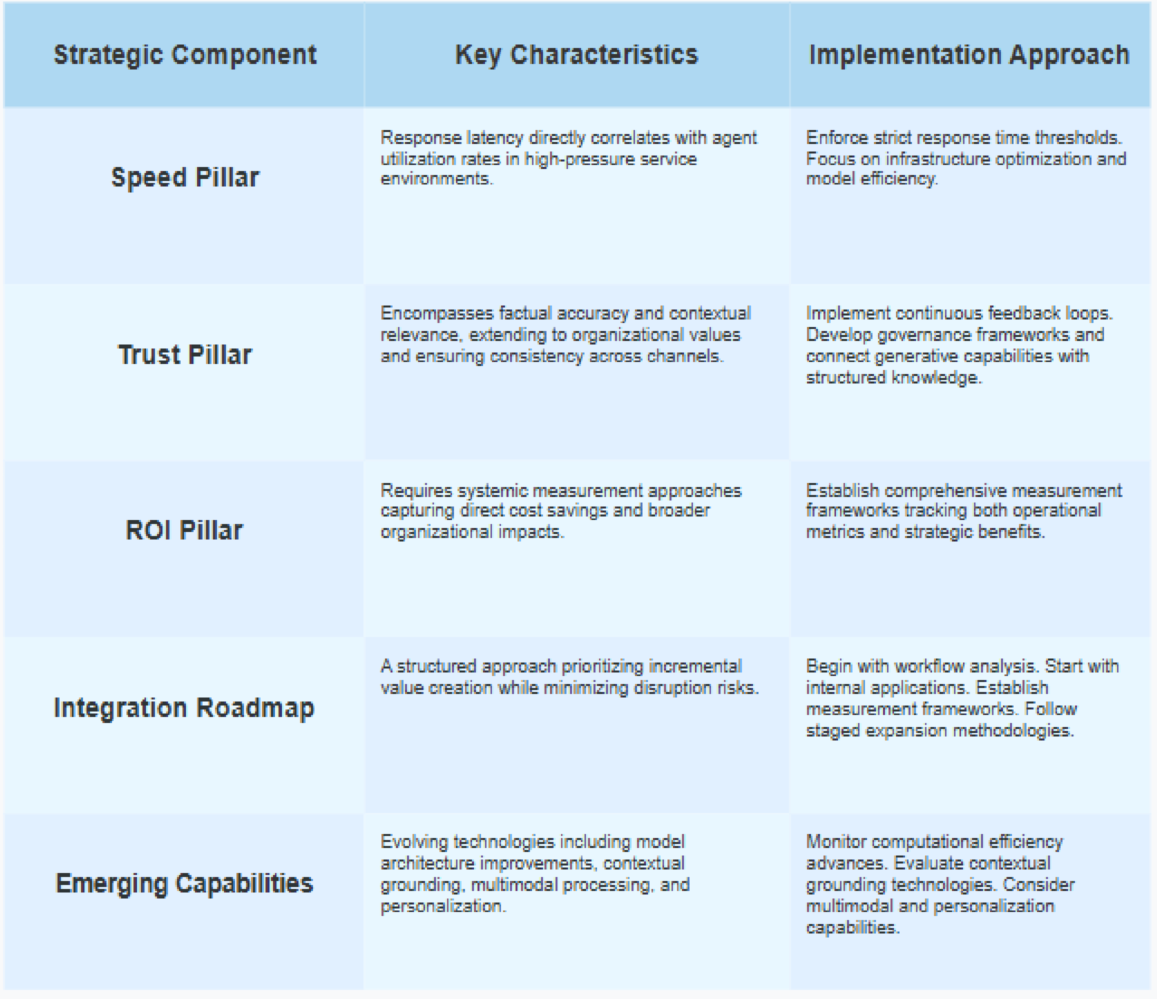
Successful GenAI implementations in customer service rest on three fundamental pillars that determine long-term viability and impact. Speed represents the foundation of adoption, with research indicating that response times directly correlate with agent utilization rates in high-pressure service environments [9]. Organizations achieving optimal latency thresholds report significantly higher engagement compared to implementations where agents experience noticeable delays. Trust emerges as the second critical pillar, encompassing both factual accuracy and contextual relevance. This pillar extends beyond mere correctness to include alignment with organizational values and established processes. Customer service environments face unique trust challenges when implementing GenAI, particularly regarding consistency across different channels and customer segments. The final pillar—ROI—requires systemic measurement approaches that capture both direct cost savings and broader organizational impacts. Industry analysis demonstrates that successful implementations maintain clear cost visibility while accurately attributing value creation across multiple dimensions [9]. The integration of these three pillars creates a balanced framework for implementation, enabling organizations to evaluate technology not merely as a cost-saving measure but as a strategic capability enhancer that transforms service operations while respecting essential human elements of customer interaction.

5.2 Integration Roadmap for Customer Service Leaders

Customer service leaders considering GenAI implementation should follow a structured integration roadmap that prioritizes incremental value creation while minimizing disruption risks. Research on successful deployments emphasizes beginning with comprehensive workflow analysis to identify high-impact opportunities where existing processes demonstrate clear inefficiencies [10]. Initial implementations show highest success rates when focusing on internal, non-customer-facing applications that allow organizations to develop capabilities and confidence before direct customer exposure. Establishing robust measurement frameworks emerges as another critical success factor, with multi-dimensional metrics providing visibility into both immediate impacts and longer-term strategic benefits. Organizations demonstrating implementation excellence consistently implement strong feedback mechanisms between frontline agents and AI systems, creating continuous improvement cycles that enhance relevance and accuracy over time. The most successful integration approaches follow staged expansion methodologies where each deployment phase builds upon validated successes from previous stages rather than pursuing comprehensive transformation simultaneously [10]. This measured approach allows organizations to optimize performance, refine use cases, and build internal advocacy while carefully managing implementation risks that could otherwise undermine confidence in the technology.

5.3 Emerging Trends and Future Capabilities

The GenAI landscape continues to evolve rapidly, with several emerging trends positioned to significantly enhance customer service capabilities. Recent innovations in model architecture and computational efficiency are addressing traditional performance limitations, enabling faster responses while reducing infrastructure requirements [9]. These advancements directly support the speed pillar while making implementations more economically viable across diverse organizational contexts. Contextual grounding technologies represent another significant development area, with emerging approaches dramatically improving factual accuracy by connecting generative capabilities with structured knowledge bases. This trend directly enhances the trust pillar by reducing potential inconsistencies while ensuring outputs align with organizational policies. Research indicates that multimodal capabilities represent a particularly promising direction for customer service applications, enabling systems to process visual, auditory, and textual information simultaneously for more comprehensive support scenarios [10]. Personalization technologies continue advancing through sophisticated customer journey integration, allowing systems to provide highly individualized support based on relationship history and preference patterns. As these capabilities mature from experimental to mainstream status, organizations maintaining focus on the three fundamental pillars will position themselves for sustained competitive advantage in increasingly complex service environments where technology augments rather than replaces human capabilities.



Strategic Component	Key Characteristics	Implementation Approach
Speed Pillar	Response latency directly correlates with agent utilization rates in high-pressure service environments.	Enforce strict response time thresholds. Focus on infrastructure optimization and model efficiency.
Trust Pillar	Encompasses factual accuracy and contextual relevance, extending to organizational values and ensuring consistency across channels.	Implement continuous feedback loops. Develop governance frameworks and connect generative capabilities with structured knowledge.
ROI Pillar	Requires systemic measurement approaches capturing direct cost savings and broader organizational impacts.	Establish comprehensive measurement frameworks tracking both operational metrics and strategic benefits.
Integration Roadmap	A structured approach prioritizing incremental value creation while minimizing disruption risks.	Begin with workflow analysis. Start with internal applications. Establish measurement frameworks. Follow staged expansion methodologies.
Emerging Capabilities	Evolving technologies including model architecture improvements, contextual grounding, multimodal processing, and personalization.	Monitor computational efficiency advances. Evaluate contextual grounding technologies. Consider multimodal and personalization capabilities.

Fig. 4: Strategic Implementation of GenAI in Customer Service [9, 10]

6. Conclusion

Generative AI represents a significant opportunity for customer service transformation, offering tangible benefits in operational efficiency, agent experience, and customer satisfaction when thoughtfully implemented. The careful integration of GPU-powered solutions enables organizations to address growing customer expectations for personalized, efficient support across multiple channels while managing operational costs effectively. Success depends on recognizing that GenAI excels as an augmentation tool rather than a replacement for human judgment, particularly in scenarios involving high-volume, pattern-rich interactions where agents previously devoted substantial time to information processing tasks. The technology demonstrates particular strength in accelerating case comprehension, suggesting responses, and surfacing relevant knowledge, allowing skilled personnel to focus on relationship management and complex decision-making. Implementation excellence requires balancing technical considerations with organizational readiness factors, including data availability, infrastructure capabilities, integration potential, governance processes, and change management strategies. Forward-looking organizations will benefit from viewing GenAI through a strategic lens that emphasizes the complementary strengths of human and machine intelligence, creating service ecosystems where technology handles routine cognitive tasks while human agents contribute empathy, judgment, and creativity. As the technology landscape continues evolving toward more efficient models with enhanced contextual understanding and multimodal capabilities, the most successful implementations will remain grounded in fundamental principles of speed, trust, and measurable business outcomes, positioning customer service as a strategic differentiator rather than merely a cost center.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Akash T, (n.d) Generative AI in customer service: Scope, adoption strategies, use cases, challenges and best practices, ZBrain. [Online]. Available: https://zbrain.ai/generative-ai-for-customer-service/#genai-use-cases-in-customer-service

[2] Ashish S, (2025) Optimizing GPU Costs for Large-Scale GenAI Inferences, Medium, 2025. [Online]. Available: https://ashish24142.medium.com/optimizing-gpu-costs-for-large-scale-genai-inference-75f4b5252b1f

[3] Deloitte Digital, (2024) Generative AI in customer service: How GenAI is transforming digital experience management, 2024. [Online]. Available: https://www.deloittedigital.com/us/en/insights/research/generative-ai-customer-service.html

[4] Dilipkumar D J, (2025) AI-Augmented Decision Making: A Framework for Enterprise Workflow Transformation, ResearchGate, 2025. [Online]. Available: https://www.researchgate.net/publication/388949003_AI-Augmented_Decision_Making_A_Framework_for_Enterprise_Workflow_Transformation

[5] Kara H, (2024) How to Use Generative AI for Enterprise Customer Service, Rasa, 2024. [Online]. Available: https://rasa.com/blog/generative-ai-for-enterprise/

[6] Melanie M, (2024) Five insights on the state of CX in 2024, CX Network, 2024. [Online]. Available: https://www.cxnetwork.com/cx-experience/articles/five-insights-on-the-state-of-cx-in-2024

[7] RSM, (2023) The 3 pillars of artificial intelligence: How AI will reshape business and the economy, 2023. [Online]. Available: https://rsmus.com/insights/economics/the-3-pillars-of-artificial-intelligence.html

[8] Sachin M, et al., (2025) How could Generative AI support and add value to non-technology companies–A qualitative study, Technovation, 2025. [Online]. Available:https://www.sciencedirect.com/science/article/abs/pii/S0166497224001743

[9] Sanjeev V, (2024) Comparative Analysis: Traditional vs. AI-Powered Customer Service Automation Software, Biz4Group, 2024. [Online]. Available: https://www.biz4group.com/blog/traditional-vs-ai-powered-customer-service-automation

[10] Xin D, et al., (2010) The Impact of Service System Design and Flow Experience on Customer Satisfaction in Online Financial Services, ResearchGate, 2010. [Online]. Available: https://www.researchgate.net/publication/247745144_The_Impact_of_Service_System_Design_and_Flow_Experience_on_Customer_Satisfaction_in_Online_Financial_Services