Leveraging Large Language Models for Real-Time Agent Assist in Contact Centers: A Framework for Reducing Average Handle Time and Improving Customer Satisfaction

Balakrishnan Devaraj
Cognizant Technology Solutions, Michigan, USA
Bala.Krish14@gmail.com

doi: https://doi.org/10.37745/ejcsit.2013/vol14n267723 Published March 31, 2026

Citation: Devaraj B. (2026) Leveraging Large Language Models for Real-Time Agent Assist in Contact Centers: A Framework for Reducing Average Handle Time and Improving Customer Satisfaction, *European Journal of Computer Science and Information Technology*, 14(2), 67-73

Abstract: *Contact centers are central nodes in enterprise customer engagement, yet they persistently contend with high operational costs and inconsistent service quality. The emergence of large language models (LLMs) presents a transformative opportunity through real-time agent assist capabilities. This paper proposes a deployment framework integrating LLMs via retrieval-augmented generation (RAG) architectures into live contact center workflows to provide agents with contextual suggestions and dynamic knowledge retrieval. Drawing on deployment observations and comparative analysis across representative enterprise environments, this study demonstrates that LLM-based agent assist systems can reduce average handle time (AHT) by 18 to 27 percent and produce measurable improvements in customer satisfaction scores. The paper discusses architectural considerations, integration challenges, ethical safeguards, and a phased adoption roadmap for practitioners.*

Keywords: large language models, contact center AI, agent assist, average handle time, customer satisfaction, RAG

INTRODUCTION

Modern contact centers handle millions of customer interactions daily across voice, chat, and digital channels. Despite significant investment in CRM platforms, interactive voice response systems, and workforce optimization tools, organizations continue to face two persistent operational challenges: elevated average handle time (AHT) and suboptimal customer satisfaction (CSAT) scores. These metrics carry direct financial consequences — unnecessary seconds of handle time, aggregated across thousands of daily interactions, represent material cost to the enterprise.

The emergence of large language models (LLMs), particularly transformer-based architectures such as those underlying GPT-4, Claude, and open-source alternatives including Llama and Mistral, has introduced substantive new possibilities for augmenting human agent capabilities in real time. Unlike traditional knowledge management systems that require agents to manually search and interpret content, LLM-powered agent assist platforms can ingest live conversation context and surface synthesized, actionable guidance within seconds. This capability fundamentally alters the information dynamics of a live service interaction and represents a meaningful departure from prior generations of contact center AI.

Deploying LLMs in production contact center environments, however, introduces non-trivial challenges that existing literature has not yet fully addressed. These include sub-second latency constraints inherent to voice interactions, hallucination risks in high-stakes service contexts, the requirement for domain-specific knowledge grounding, and significant integration complexity when connecting AI inference pipelines with legacy telephony and CRM infrastructure. This paper addresses these challenges by presenting a practical, enterprise-grade deployment framework informed by real-world implementation observations.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3 presents the proposed framework and methodology. Section 4 reports results and findings. Section 5 discusses the implications. Section 6 addresses implications for research and practice. Section 7 concludes, and Section 8 identifies directions for future research.

LITERATURE AND THEORETICAL UNDERPINNING

Evolution of Contact Center Technology

The technological history of contact centers spans several distinct generations of innovation. Early call centers of the 1970s and 1980s relied entirely on human agents guided by paper scripts and supervisor oversight. The widespread adoption of customer relationship management (CRM) systems during the 1990s introduced structured data management, allowing agents to access customer history during interactions. Knowledge management systems in the 2000s provided searchable repositories of policies, procedures, and troubleshooting guides, though navigation remained manual and time-consuming.

Rule-based chatbots and early IVR enhancements in the 2010s began automating a subset of routine inquiries. Academic literature from this period (Zeithaml, Bitner & Gremler, 2006; Bitner, Ostrom & Meuter, 2010) documented the growing importance of technology-enabled service quality and the inherent tension between automation efficiency and personalized customer experience. These contributions established the theoretical grounding for understanding that poorly designed automation can reduce rather than enhance satisfaction outcomes.

Emergence of Conversational AI

The introduction of neural network-based natural language processing systems around 2016–2018 marked a transition point. Systems such as Google Dialogflow and IBM Watson Assistant demonstrated that intent recognition and entity extraction could achieve production-grade accuracy within constrained interaction domains. Jurafsky and Martin (2019) and contributions from the open-source development community advanced understanding of multi-turn dialogue management, slot filling, and contextual state tracking.

These first-generation conversational AI systems proved brittle in practice. They required extensive intent taxonomies, struggled with paraphrastic variation, and lacked the ability to reason over long conversational contexts. Their deployment as agent-facing assist tools yielded limited AHT improvements, as agents frequently found suggestions either too generic or contextually misaligned with the live interaction (Folstad & Skjuve, 2019).

Large Language Models and Service Applications

The release of GPT-3 in 2020 and subsequent model generations altered the landscape fundamentally. LLMs trained on large corpora demonstrated emergent capabilities in instruction-following, contextual reasoning, and generative summarization that substantially exceeded prior NLP architectures. Brown et al. (2020) documented few-shot learning capabilities enabling rapid domain adaptation without extensive fine-tuning, a property with significant implications for enterprise deployment economics.

Subsequent work on retrieval-augmented generation (RAG) by Lewis et al. (2020) addressed a critical limitation — the tendency of LLMs to generate factually unsupported content. By grounding model outputs in retrieved document passages, RAG architectures demonstrated substantially improved factual accuracy in enterprise knowledge retrieval tasks. Shuster et al. (2021) further validated that retrieval augmentation materially reduces conversational hallucination rates, making the architecture considerably more appropriate for high-stakes contact center environments.

Theoretical Framework

This study is grounded in two complementary theoretical perspectives. First, cognitive load theory (Sweller, 1988; Wickens & Hollands, 2000) provides a framework for understanding how reducing extraneous cognitive burden on agents — specifically the burden of real-time knowledge retrieval — can improve both task performance and interaction quality. Second, service quality theory (Berry, 1995; Zeithaml et al., 2006) informs the proposition that technology-mediated reduction in agent cognitive load translates to perceivable improvements in service encounter quality from the customer perspective.

METHODOLOGY

Framework Design

The proposed framework, designated the Contextual Agent Assist Layer (CAAL), operates as a middleware intelligence layer positioned between the contact center's interaction platform and its underlying knowledge systems. CAAL continuously monitors live interaction streams, constructs a dynamic context representation, and generates agent-facing suggestions through a RAG pipeline. The framework comprises five integrated components, described in Table 1 below.

Table 1. CAAL Framework Component Overview. The five components operate as an integrated pipeline from raw interaction stream to agent-facing suggestion.

Component	Function	Key Technology
Transcription Engine	Converts voice/text streams to structured text with <250ms latency	ASR models (Whisper, Deepgram); WebSocket streaming
Context Aggregation Module	Maintains rolling conversation window with CRM data and intent signals	Event-driven architecture; CRM API connectors; intent classifiers
RAG Retrieval Engine	Semantic search across enterprise knowledge bases for relevant passages	Vector database (Pinecone, Weaviate); dense passage retrieval
LLM Synthesis Layer	Generates concise suggestions by combining retrieved passages with conversation context	Fine-tuned LLM with domain adaptation; confidence scoring; response filtering
Agent Interface Renderer	Presents suggestions in agent desktop with accept, edit, or dismiss controls	CRM-embedded widget; REST API integration; audit logging

Data Collection Approach

To evaluate the operational impact of CAAL, this study draws on implementation observations from two enterprise contact center environments: one in the financial services sector and one in the telecommunications industry, each operating with more than 300 active agents. Performance data was collected over a 90-day pre-implementation baseline period and compared against a 90-day post-implementation period following a structured, phased rollout.

Four primary metrics were tracked: average handle time (AHT) in seconds; customer satisfaction score (CSAT) collected through post-interaction surveys on a ten-point scale; first contact resolution (FCR) rate expressed as a percentage of interactions resolved without escalation or callback; and agent task load index (TLX) scores adapted from the NASA cognitive workload instrument for contact center conditions. Operational variables including call volume seasonality, staffing changes, and product line updates were monitored throughout the observation period and are accounted for in the interpretation of findings.

Phased Deployment Protocol

Both environments followed a structured three-phase deployment protocol. Phase 1 (Weeks 1–8) deployed CAAL to a pilot cohort of 20 to 30 agents handling a single defined interaction type, focusing on system stability, latency benchmarking, and early friction identification. Phase 2 (Weeks 9–20) extended deployment to additional teams and interaction types, introducing continuous model feedback loops capturing agent accept and dismiss signals for retrieval ranking refinement. Phase 3 (Weeks 21 onward) achieved full deployment with ongoing performance monitoring, A/B testing of suggestion generation strategies, and expanded knowledge base coverage.

RESULTS AND FINDINGS

Average Handle Time

Across both environments, mean AHT decreased substantially following CAAL implementation. In the financial services environment, AHT decreased from a baseline mean of 487 seconds to 371 seconds, representing a 23.8 percent reduction. In the telecommunications environment, AHT decreased from 412 seconds to 336 seconds, a reduction of 18.4 percent. These figures are consistent with the range projected in the framework design phase and align with outcomes reported in comparable industry practitioner deployments.

Analysis of interaction-level data identified two primary sources of time savings. First, knowledge retrieval tasks, in which agents previously spent 45 to 90 seconds navigating internal knowledge bases to locate policy or procedure information, were substantially accelerated by CAAL's real-time retrieval suggestions. Second, wrap-up tasks were reduced by an estimated 30 to 40 percent through LLM-generated post-interaction summaries, restructuring the cognitive workflow both during and after the interaction.

Customer Satisfaction

Post-interaction CSAT surveys indicated meaningful improvement across both environments. The financial services environment recorded mean CSAT improvement from 7.1 to 8.3 on the ten-point scale, while the telecommunications environment improved from 6.8 to 7.9. Open-text survey responses and agent debrief sessions attributed these improvements to three themes: reduced wait time while agents located information; greater accuracy and consistency of information provided; and a perception that agents were more attentive during the interaction, reflecting the cognitive load reduction effect predicted by theory.

First Contact Resolution and Agent Task Load

First contact resolution rates improved by approximately 15 percent in both environments, from a combined baseline average of approximately 70 percent to approximately 80 percent post-implementation. This improvement is primarily attributed to the assist system surfacing complete resolution pathways — including edge cases and escalation decision trees — that agents might otherwise have missed or imperfectly executed under time pressure.

Agent TLX scores decreased by approximately 18 to 19 percent across environments. Qualitative feedback from structured group discussions indicated that agents who perceived the system as a cognitive partner rather than a surveillance mechanism demonstrated significantly higher utilization and acceptance rates. Environments in which the system was framed as a performance monitoring tool experienced elevated agent resistance and utilization rates plateauing at approximately 40 to 50 percent.

Table 2. Pre- and Post-CAAL Implementation Performance Metrics. FS = Financial Services; Telecom = Telecommunications. Percentage figures represent change relative to baseline.

Metric	FS Baseline	FS Post-CAAL	Telecom Baseline	Telecom Post-CAAL
Avg. Handle Time (seconds)	487	371 (-23.8%)	412	336 (-18.4%)
CSAT Score (/10)	7.1	8.3 (+16.9%)	6.8	7.9 (+16.2%)
First Contact Resolution (%)	71.4%	82.1% (+15.0%)	68.9%	79.3% (+15.1%)
Agent TLX Score (lower = better)	64.2	51.8 (-19.3%)	67.1	55.4 (-17.4%)

DISCUSSION

Interpretation of AHT and CSAT Findings

The AHT reductions observed across both environments — 18.4 and 23.8 percent respectively — exceed the typical 10 to 15 percent improvements reported in earlier deployments of rule-based assist systems, suggesting that LLM-based contextual synthesis delivers efficiency gains that prior approaches could not achieve. The simultaneous improvement in CSAT scores is theoretically coherent with the cognitive load framework underpinning the study: agents whose extraneous cognitive burden is reduced are better positioned to maintain attentive, empathetic interaction with customers, producing service encounter experiences that customers rate more favorably.

The correlation between these outcomes challenges the conventional wisdom that speed and quality in contact center service represent a tradeoff. The CAAL data suggests that, at least in the context of well-implemented LLM assist, improvements in handling efficiency and improvements in service quality can be achieved concurrently. This finding has significant strategic implications for contact center leaders who have historically managed AHT and CSAT as competing priorities.

Infrastructure and Integration Considerations

Achieving the latency targets required for voice-channel agent assist — end-to-end pipeline completion within 1,500 to 2,000 milliseconds — required co-locating inference infrastructure with contact center application servers, applying model serving optimizations including quantization and speculative decoding, and engineering vector search performance below 100 milliseconds. Organizations underestimating infrastructure complexity will encounter latency degradation that impairs agent trust in the system and reduces utilization rates.

Knowledge base preparation consistently emerged as an underestimated component of deployment readiness. Enterprise knowledge artifacts require chunking strategies, metadata tagging, and de-duplication before effective semantic retrieval is achievable. Both deployment environments budgeted between three and four months of knowledge engineering effort prior to system launch, a timeline that should be incorporated into project planning for prospective deployments.

Ethical and Governance Dimensions

Real-time transcription of interactions for LLM inference raises data governance obligations under applicable privacy frameworks including GDPR and CCPA. Organizations must ensure appropriate legal bases for processing, customer notification compliance, and contractual data processing agreements with AI vendors. The distinction between inference-only processing and model training data usage requires explicit internal policy codification.

In regulated industries, the risk of LLM-generated suggestions being factually inaccurate or non-compliant represents a regulatory liability beyond mere quality concern. The CAAL framework addresses this through confidence-based filtering and the grounding of all suggestions in retrieved, audited knowledge artifacts. Nonetheless, comprehensive audit logging of all generated suggestions and agent response actions is a non-negotiable governance requirement for deployments in regulated contexts.

Implication to Research and Practice

Implications for Research

This study contributes to the emerging literature on applied LLM deployment in enterprise service environments by presenting a complete architectural framework and associated performance evidence from production deployments. The simultaneous improvement in both AHT and CSAT observed in this study challenges prior theoretical models that treat these metrics as inherently opposed, and invites further investigation into the mechanisms through which cognitive load reduction produces service quality effects perceivable to customers.

The study also surfaces methodological considerations for future research in this area. The absence of randomized controlled trial conditions — an inherent limitation of applied deployment studies — necessitates care in causal attribution. Future work employing experimental or quasi-experimental designs, potentially through staged rollouts with matched agent cohorts, would strengthen the evidence base. Additionally, longitudinal studies examining whether initial AHT improvements are sustained or diminished as agent familiarity with the system matures would address an important open question.

Implications for Practice

For contact center practitioners, this study offers a deployment framework that is directly actionable and grounded in production experience. The three-phase adoption roadmap, the knowledge engineering timeline estimates, and the change management principles derived from observational findings provide implementation guidance that is not available from theoretical literature alone. Practitioners should particularly attend to the framing of the system to agents and supervisors, as this factor emerged as a significant determinant of utilization and therefore of realized operational benefit.

For technology leaders evaluating LLM vendors, the framework's emphasis on RAG architecture as a prerequisite for hallucination risk management offers a practical procurement criterion. General-purpose LLMs deployed without retrieval grounding present unacceptable accuracy risks for most contact center use cases, particularly in regulated industries. The selection of vector database infrastructure capable of sub-100-millisecond search performance at scale should likewise be treated as a baseline technical requirement rather than an optimization.

CONCLUSION

This paper has presented the Contextual Agent Assist Layer (CAAL), a structured framework for deploying large language models as real-time assist tools in enterprise contact center environments. Grounded in retrieval-augmented generation principles, cognitive load theory, and service quality research, the framework addresses the principal technical, organizational, and ethical challenges that have constrained LLM-based agent assist adoption at scale.

Performance observations from two enterprise deployments demonstrate that well-implemented LLM agent assist can reduce AHT by 18 to 24 percent while simultaneously improving CSAT scores by approximately 16 percent and first contact resolution rates by approximately 15 percent. These outcomes are contingent on deliberate infrastructure planning, sustained knowledge engineering investment, and change management practices that position the system as an agent empowerment tool rather than a performance surveillance mechanism.

The convergence of advancing LLM inference efficiency, declining AI infrastructure costs, and maturing RAG architectures suggests that real-time agent assist will become a standard component of enterprise contact center architecture within the near term. Organizations that develop deployment capabilities and governance frameworks now will be better positioned to extract competitive advantage from this transition.

Future Research

Several directions merit investigation beyond the scope of the present study. First, longitudinal analysis examining whether AHT improvements are sustained over periods beyond the initial 90-day post-implementation window would address the question of novelty effects versus durable behavioral change. Second, comparative studies of different LLM architectures — including smaller, domain-specific models against large general-purpose models — under real-time latency constraints would provide actionable infrastructure guidance for organizations at various deployment scales.

Third, research into optimal RAG chunking and hybrid sparse-dense retrieval strategies specifically calibrated for contact center knowledge artifacts — characterized by procedural, structured content rather than expository prose — could yield further efficiency gains in suggestion relevance. Fourth, investigations into the equity implications of LLM assist deployment, examining whether suggestion quality is systematically differentiated across customer demographic segments due to ASR accuracy or training data bias, represent an important ethical research frontier. Finally, the extension of LLM assist frameworks to post-interaction workflows, including automated quality assurance and coaching recommendation generation, presents a natural and high-value expansion of the applications examined in this paper.

REFERENCES

Berry, L. L. (1995). Relationship marketing of services — growing interest, emerging perspectives. Journal of the Academy of Marketing Science, 23(4), 236–245.

Bitner, M. J., Ostrom, A. L., and Meuter, M. L. (2010). Implementing successful self-service technologies. Academy of Management Executive, 16(4), 96–109.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Folstad, A. and Skjuve, M. (2019). Chatbots for customer service: User experience and motivation. Proceedings of the 1st International Conference on Conversational User Interfaces, 1–9.

Jurafsky, D. and Martin, J. H. (2019). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Pearson.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Nakano, R., Hilton, J., Balwit, A., Wu, J., Ouyang, L., Kim, C., and Schulman, J. (2022). WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. Findings of the Association for Computational Linguistics: EMNLP 2021, 3784–3803.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285.

Wickens, C. D. and Hollands, J. G. (2000). Engineering psychology and human performance (3rd ed.). Prentice Hall.

Zeithaml, V. A., Bitner, M. J., and Gremler, D. D. (2006). Services marketing: Integrating customer focus across the firm (4th ed.). McGraw-Hill.