White Paper

intel ai

Retail, Banking/Financial Services, Health Care

Celebal Helps Call Centers Keep Up with Real-Time Demands

Celebal's AICXM makes use of the CPU, GPU and NPU built into Intel AI PC-based systems to provide real-time insight to customer conversations combined with advanced post-call analytics

Authors

Shailesh Pandey
AI Solution Architect

Marcel Wagner
Principal Engineer

Yatindra Shashi
AI cloud Engineer

Harsh Madaan
Global Head, AICXM

Balan Dhanka
Senior AI Advisor, AICXM

Sagar Sarolia
AI Scientist, AICXM

Table of Contents

Problem Statement 1
Celebal AICXM Powered by Intel AI PC 2
Using AI to Improve Call Center Quality 2
AICXM Technology Stack 2
Intel AI PC Delivers Hybrid Compute Performance 2
Optimizing NPU Workloads with OpenVINO™ Toolkit 2
Live Call Analysis: Addressing Customer Calls in Real Time 3
Call Insights at Scale 4
Demonstrating AICXM Performance 5
PCA Performance 5
Conclusion 7
Learn More 7

Customers are more educated and demanding about products thanks to content creators, YouTube videos, product forums and other internet information resources. This makes the job of a call center agent more difficult and puts pressure on companies to maintain a great call center customer experience to ensure customer satisfaction, retention, and brand reputation.

On the frontlines, supervisors need more data on live agent interactions but find it time-consuming and costly to sit in on a large number of calls. Without this information, it is a challenge to coach agents in the moment or course-correct conversations before they go off-track, and calls can't be analyzed at scale.

This customer experience upleveling is on top of ongoing demands to be compliant with company and governmental mandates.

Problem Statement

Today's contact centers need an approach that matches the speed and complexity of modern customer expectations. Traditional methods of monitoring calls, manually reviewing performance, and relying on siloed systems are not scalable or efficient. Enterprises face challenges in:

- Handling increasingly complex customer queries.
- Providing supervisors with real-time visibility into live calls.
- Coaching agents effectively during ongoing interactions.
- Ensuring compliance with regulatory and organizational standards.

These limitations increase operational costs, reduce customer satisfaction, and impact long-term brand loyalty.

The objective of this paper is to explore how AI-powered, on-premises solutions can modernize customer experience management in contact centers. Specifically, this paper highlights how Celebal Technologies' Artificial Intelligence Customer Experience Management (AICXM), when deployed on Intel AI PC architecture, addresses current industry pain points by providing:

- Real-time automation and visibility into calls.
- Intelligent agent assistance during live interactions.
- Scalable, privacy-first deployment without dependency on cloud infrastructure.
- Advanced analytics for both live call analysis (LCA) and post-call analysis (PCA).

Celebal AICXM Powered by Intel AI PC

That's why Celebal Technologies, an Intel® Industry Solutions Builders member, created Artificial Intelligence Customer Experience Management (AICXM), an intelligent conversation platform designed to bring real-time automation, visibility, and intelligence to every call.

AICXM empowers agents with AI-driven prompts, retrieves relevant information mid-conversation, and enables supervisors to analyze performance instantly through searchable transcripts and live sentiment tracking.

Using AI to Improve Call Center Quality

AICXM is an on-premises AI-enabled solution that modernizes and elevates the customer experience in contact center environments. Developed as part of Celebal's Intelligent Customer Experience (ICX) portfolio, AICXM integrates advanced AI models, Whisper speech transcription, and orchestration technologies to deliver real-time agent assistance and post-call analytics—without the need for data migration or cloud dependency.

AICXM has the ability to unify and analyze customer interactions in real time, even when data is dispersed across multiple, siloed systems. Contact center agents often navigate two or more display screens simultaneously, juggling customer queries while switching between screens and databases.

AICXM Technology Stack

The AICXM is deployed on AI PC-based systems powered by Intel, offering a privacy-first, low-latency alternative for enterprises with strict data governance and performance requirements.

AICXM features a modular microservices architecture running on K3s—a lightweight Kubernetes distribution well-suited for on-premises environments. This makes the system easy to scale and integrate with existing infrastructure without requiring the software to be hosted on a cloud server.

The solution uses the Llama 3.2 AI large language model (LLM) that is optimized for scalability when run on the K3s microservice architecture.

One of the important technologies used in the AICXM is the Open Platform for Enterprise AI (OPEA) which provides functionality building blocks that allow the development of an AI workflow that is holistic and overcomes implementation fragmentation.

OPEA works with the LLM to ensure a scalable solution that can serve many users. Key OPEA function blocks include:

- Ingest/data processing
- Embedding models/services
- Indexing/vector/graph data stores
- Retrieval/ranking
- Prompt engines
- Guardrails: Llama Guard was used for PCA and LCA
- Memory systems

In the AICXM, OPEA is used to create a retrieval augmented generation (RAG) workflow allowing Llama to ingest data from an enterprise or external repository and insert it into a workflow. RAG plays an important role in the live call analysis (LCA) and post-call analysis (PCA) which are described below.

The tests used Llama Guard for moderation, deployed as an OPEA-provided Kubernetes microservice. It runs both pre-LLM (on user inputs) and post-LLM (on model outputs) to classify content as safe/unsafe and tag violations by predefined categories. In the test setup, it was deployed on the same GPU-serving infrastructure as the main OPEA Framework.

Intel AI PC Delivers Hybrid Compute Performance

AICXM is purpose-built to take full advantage of Intel AI PC architecture, delivering powerful, real-time AI performance. Powered by Intel® Core™ Ultra Processors (Series 2), the solution benefits from the hybrid compute capability of integrated CPU, GPU, and NPU.

AICXM leverages the 16-core CPU of the Intel Core Ultra Processor to maintain smooth operation of its microservice infrastructure and to ensure responsive interaction with agent and supervisor dashboards.

The integrated GPU in the Intel® Core™ Ultra platform handles the compute-intensive AI operations, including execution of the Llama 3.2 model. This model is used in both LCA and PCA contexts for generating summaries, extracting KPIs, and interpreting customer-agent interactions in natural language.

To accelerate Llama 3.2 inferencing, the AICXM uses IPEX LLM, a PyTorch library optimized for Intel® processors. IPEX automatically optimizes execution across a broad range of Intel® architecture processors (CPUs, GPUs, and NPUs) using the best hardware for each part of the workload.

The Intel® GPU Device Plugin is used to facilitate Kubernetes workload offloading to the GPU. The GPU's highly parallel architecture ensures efficient LLM inference, making it ideal for secure, on-device natural language processing that is responsive and scalable.

Complementing the GPU is the Intel® Neural Processing Unit (Intel® NPU), a specialized compute engine designed to accelerate AI inference tasks with minimal power draw. In an AICXM application, the NPU runs the Whisper model for speech-to-text conversion during both LCA and PCA operations. This offloads transcription from the CPU and GPU, improving overall system efficiency while maintaining real-time responsiveness.

Optimizing NPU Workloads with OpenVINO™ Toolkit

The AICXM integrates OpenVINO™ toolkit for optimizing Whisper workloads running on the NPU. By running Whisper on the NPU through OpenVINO toolkit, the solution is able to maximize inference speed and reduce latency during real-time transcription. OpenVINO toolkit also improves model portability and deployment consistency across diverse Intel® hardware in edge environments, making it easier for integrators to deploy AICXM at scale.

Live Call Analysis: Addressing Customer Calls in Real Time

AICXM contains two main functional components: Live Call Analysis (LCA) and Post Call Analysis (PCA). These capabilities enable edge AI integrators to address both the real-time demands of active customer support and the retrospective needs of performance monitoring, compliance, and coaching.

LCA (see Figure 1) supports active call sessions via the following workflow:

- Call Initiation
  - Users connect via Twilio; audio is streamed in real-time to both the AI inference pipeline and the agent desktop through the LCA Portal UI.
- Real-Time Transcription
  - Audio is transcribed on-device using Whisper, executed on the NPU for ultra-low latency using Safe Streams Transport (SST).
- Orchestration and Pipeline Management via OPEA
  - The OPEA framework, deployed as a Kubernetes microservice, coordinates the entire pipeline—from transcription to retrieval to response generation—maintaining state and context across asynchronous data streams.
- RAG Retrieval with Redis Vector DB
  - Transcribed text is passed to the RAG module.
  - bge-base embeddings are computed on single-thread CPU with a batch size of 1,500 tokens.
  - Similar documents are retrieved from the Redis 7.2 vector database.
  - A reranker model filters top results for semantic fit. On typical consumer GPUs, the open-source BGE reranker is fast enough for production top-k reranking, with the lightweight bge-reranker-v2-m3 often delivering sub-100–200 ms per query for small k on a single consumer-class GPU; exact latency depends on sequence length, batch size, and k, but it is notably quicker than larger cross-encoders and slower than the fastest hosted APIs.
- LLM Inference with Llama 3.2 on the GPU
  - The reranked context and user intent are sent to Llama 3.2, running in a container on the GPU.
  - The Intel GPU Device Plugin for Kubernetes ensures the Llama workload is dynamically scheduled to the integrated GPU, enabling isolated, GPU-accelerated inference within K3s-managed pods.
- Suggested Response Delivery
  - Llama's response is surfaced to the call center agent in real time through the LCA Portal, supporting human-in-the-loop decision-making and improving call resolution rates.

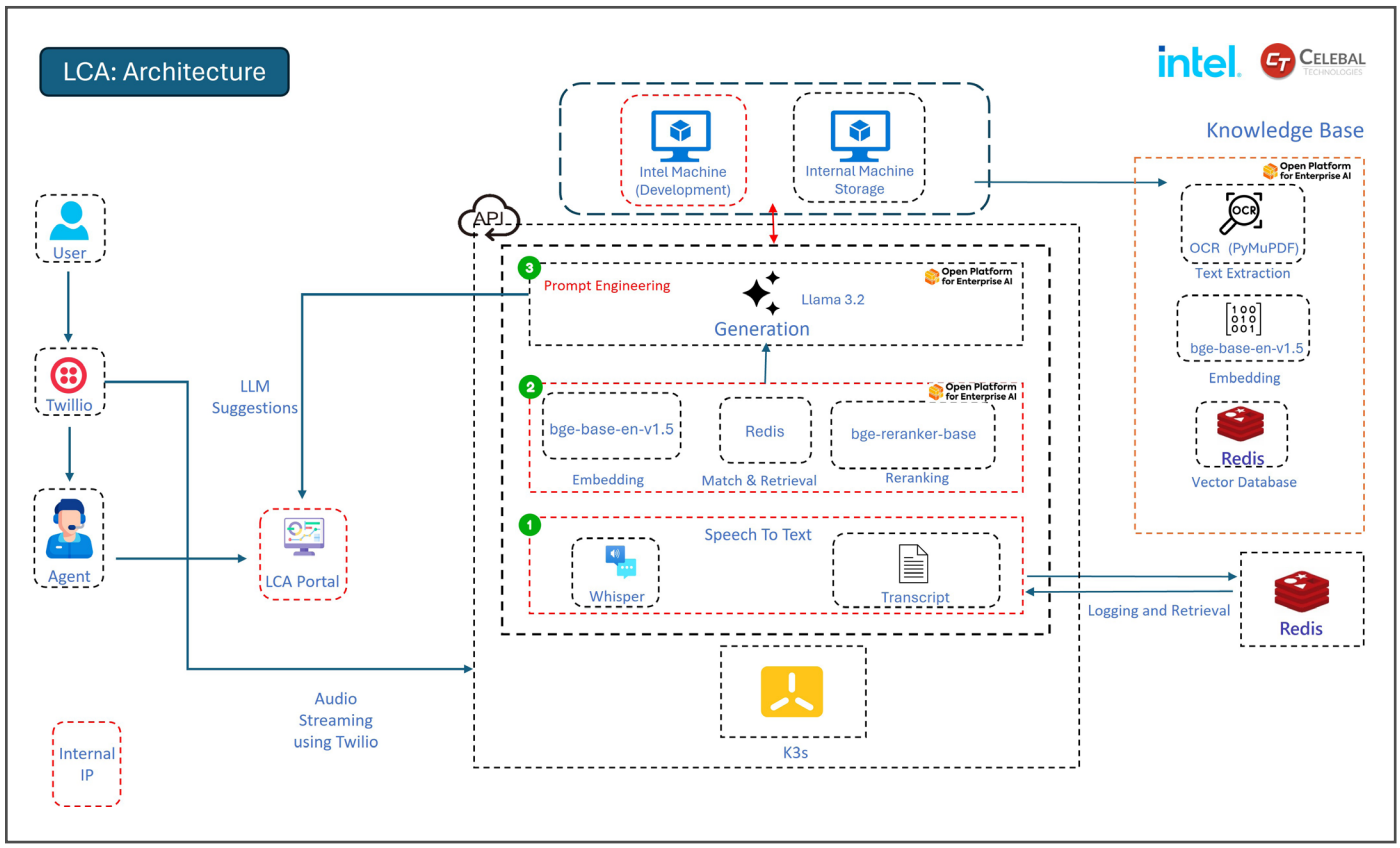

Figure 1. LCA architecture and data flow.

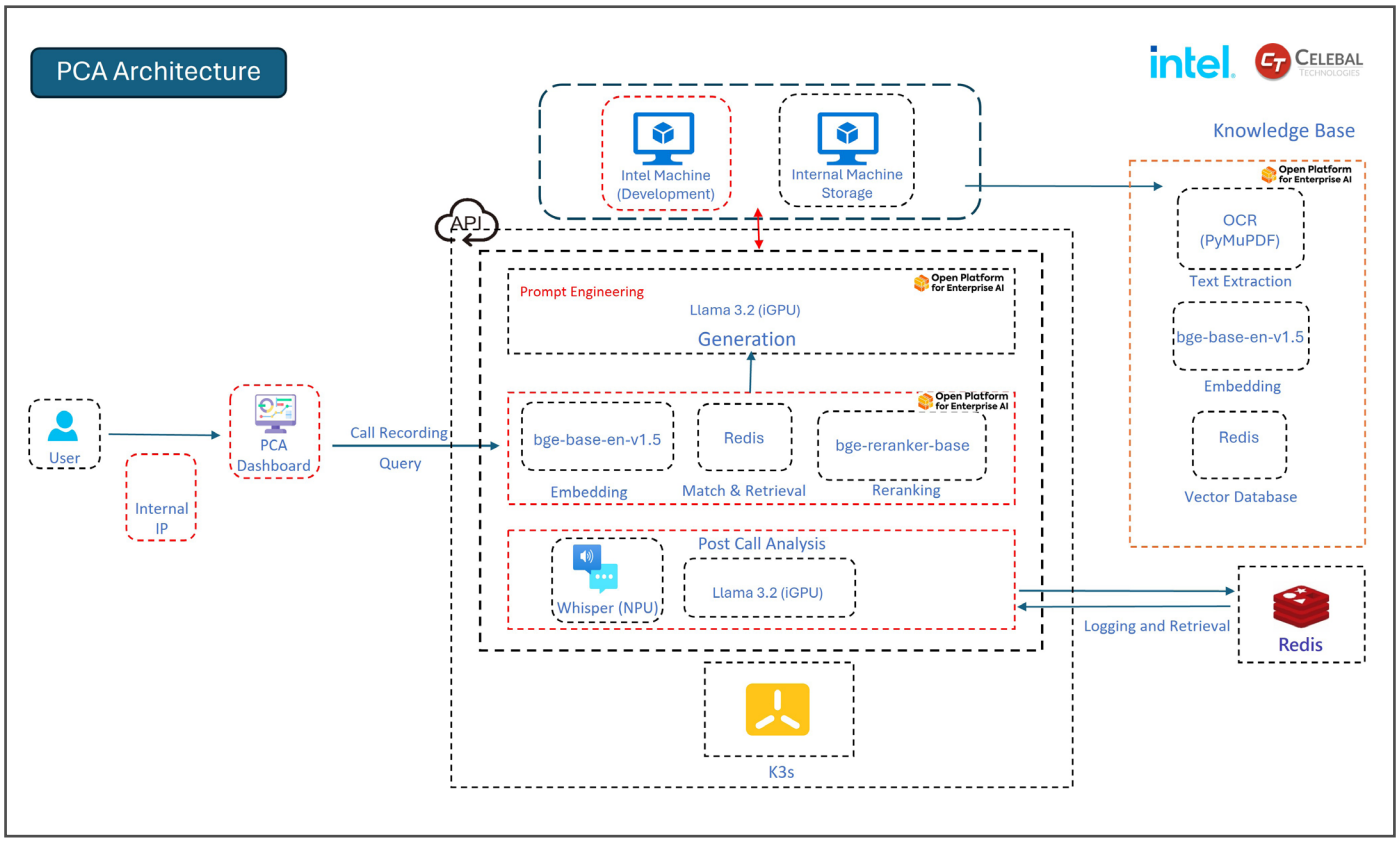

Figure 2. PCA architecture and data flow.

The LCA pipeline's ability to parse conversation context, pull in relevant knowledge from multiple sources, and deliver intelligent recommendations without delay translates into measurable business benefits. The LCA enables enterprises to improve first call resolution (FCR), reduce call handling time, and increase overall customer satisfaction. Because the system is deployed entirely on premises, it also ensures real-time compliance monitoring—especially important for industries such as banking, financial services, and insurance, where sensitive data and strict regulatory requirements are standard.

Call Insights at Scale

The job of the PCA is to derive insights from completed interactions (see PCA architecture in Figure 2).

Once the live call is over, users can access the following user interaction and input information from the PCA dashboard over a secure intranet. The dashboard includes:

Archived call recordings

- User queries related to call intent, compliance, or customer satisfaction
- Microservice-orchestrated processing pipeline (via K3s + OPEA)
- The entire PCA stack is containerized using K3s and managed through the OPEA framework, which coordinates model flow and data access in real time.

Pipeline steps:

- Transcription: Whisper-Medium-en transcribes archived audio, leveraging the NPU for low-power INT8 inference.
- Document Parsing: OCR is performed using Tesseract optical character recognition engine on local files, managed by CPU containers.
- Embedding Generation: bge-base-en-v1.5 converts both documents and queries into vector embeddings, processed on the CPU.

RAG workflow with Redis + Reranking:

- Embeddings are stored in Redis 7.2 vector database.
- Query embeddings are matched against document vectors.
- Top-k candidates are reranked using bge-reranker-base.
- Final prompt is constructed using engineered templates tailored to PCA use cases.

LLM inference with Llama 3.2 on the GPU via Intel GPU Device Plugin

- Llama 3.2 runs inference using retrieved context, containerized with K3s.
- The Intel GPU Device Plugin for Kubernetes ensures iGPU access is exposed as a schedulable resource to K3s pods, allowing dynamic placement of Llama containers onto the iGPU.

- This hardware-aware plugin provides isolation, parallelism, and optimal GPU utilization, crucial for latency-sensitive PCA pipelines.

Final output and KPI extraction

The generated response is parsed for:

- Customer satisfaction (CSAT) summaries
- Agent performance metrics
- Compliance flags and anomalies

These are then visualized via the PCA dashboard, enabling supervisors to extract actionable insights without manual review.

The PCA provides a complete record of the customer's experience and the insights that can be gained from that interaction, as well as trends from all of the customer interactions.

Demonstrating AICXM Performance

To show its performance, Celebal put the AICXM through a series of tests[i]. Hardware for the tests include Intel® Core™ Ultra processors client platform featuring a 16-core CPU, Intel® Arc™ B-series GPU and a fourth-generation NPU. The platform supported 32GB of RAM with 1TB of SSD memory.

Testing of software performance was done using Retrieval Augmented Generation Assessment System (RAGAS). Applications being tested included Whisper running in INT8 mode, Llama, and Redis vector database. Core software on the platform included OpenVINO toolkit, PyTorch, K3s and the PyMuPDF Python library (Fitz).

In these tests the AI PC-based solution experienced 230 ms of inference latency. Power consumption for the AI PC version was 15W peak.[1]

PCA Performance

Two tests were run on post call analytics (PCA). Figure 3 shows the average RAGAS score from 75 test rounds of calls.

The performance of answer relevancy was high with a 0.81 which shows that AICXM can closely match user queries. The faithfulness score of 0.78 indicates that the system offers truthful and consistent information whereas the context precision mark also demonstrates precise understanding of relevant content.[1] The context recall score shows the system is adept at retrieval of contextually relevant information.

The harmfulness score was zero and should always be zero since the generation of unsafe or harmful outputs, even at low frequency, constitutes a critical failure that compromises model alignment, violates responsible AI principles, and undermines robustness, trustworthiness, and downstream applicability.

Figure 4 shows the results of the PCA for eight audio clips that represented actual phone calls. The duration of the call (dark blue) ranges from 1.45 minutes to 5.18 minutes. The tests then measured the time of a workflow that included transcription (light blue), word error rate (yellow), extraction of KPI data (red) and time needed to generate a summary.

[1] Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

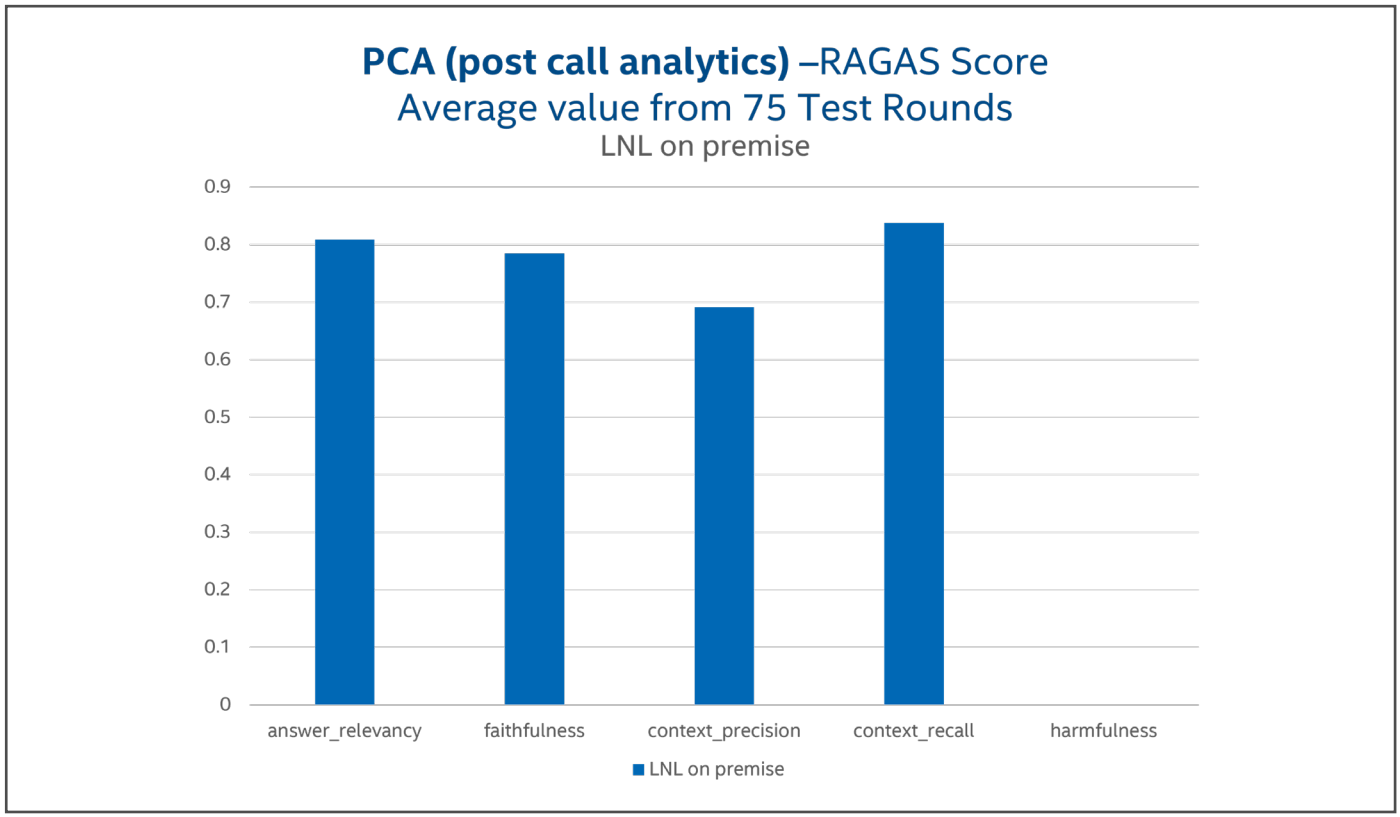

Figure 3. Average RAGAS scores showing performance and quality of the AICXM (higher is better). For harmfulness, lower is better.

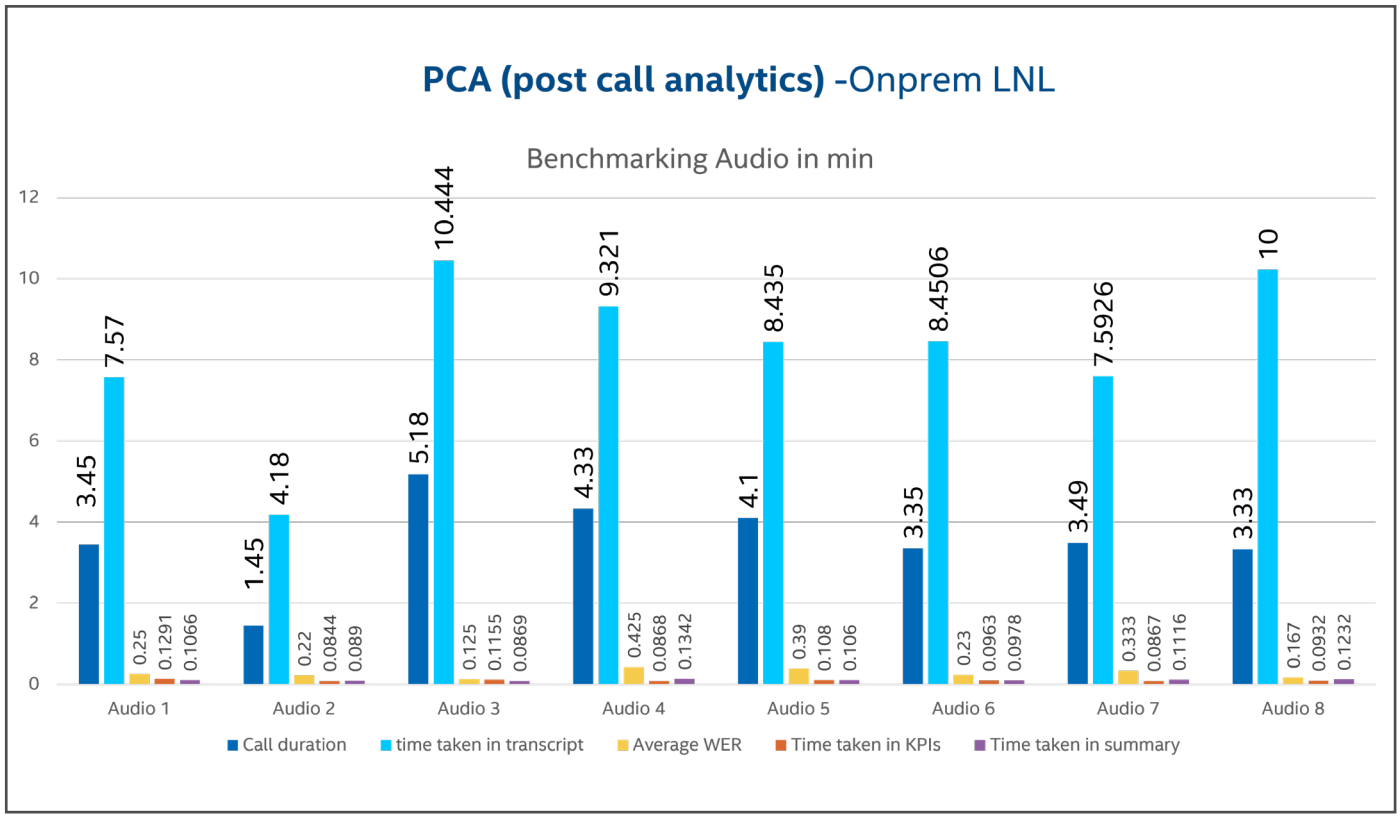

Figure 4. PCA performance for five key performance analytics.

The results show that transcription time is consistently between two and three times longer than call duration, ranging from ~4.18 to 10.44 minutes.

LCA achieves real-time transcription by processing audio in small chunks through a streaming pipeline, whereas PCA uses batch processing over full recordings, which is more compute-intensive. On-prem compute restrictions further slow PCA, making it run 2–3× longer than call duration, while LCA remains optimized for low-latency output.

Average Word Error Rate (WER) varies between 0.12 and 0.43, indicating transcription accuracy. The efficiency of the system's KPI extraction and summary generation is shown in these tests where they consistently remain under 0:15 minutes.[2]

Overall, the tests show effective transcription and analytics performance with manageable latency.

[2] Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Conclusion

The Celebal AICXM is the answer for call centers in a wide range of industries that need to better support their agents with insight and data when they need it most – while they are engaged with the customer. It offers the AI-enabled capability to analyze calls and find relevant documents and scripts so the agent can better satisfy the customer. To enable this solution, Celebal turned to Intel's AI PC product architecture to offer the performance needed for its solution.

This close collaboration with Intel enables Celebal to deliver a tightly optimized, fully local AI platform that combines the performance of modern AI models with the privacy and control of on-premises deployment. It's a solution built to meet the evolving needs of enterprise customer experience teams—delivering smarter interactions, faster responses, and deeper insight without compromise.

Learn More

Celebal Technologies

AICXM by Celebal Technologies

Open Platform for Enterprise AI (OPEA)

IPEX LLM

Intel® Core™ Ultra Processors

AI PCs Powered by Intel

Intel® Arc™ Graphics

OpenVINO™ Toolkit

Intel® GPU Device Plugin

Device Plugins: The Path to Faster Workloads in Kubernetes

Intel® Industry Solution Builders

intel ai

[i] **SUT**: 1-node, 1x Intel Core Ultra processor 9 288V with integrated NPU and Arc series 16GB iGPU. Total DDR5 memory was 32 GB (2 slots/16 GB/5600 MHz); Intel® Hyper-Threading Technology – enabled; Intel® Turbo Boost Technology – enabled. Storage: 512 GB

Software: OS was Windows 11 using WSL2.

Benchmark/workload software: CoCIF (Call Center Intelligence Framework) pipeline including speech-to-text and post-call analytics.

Libraries: OpenVINO™ Toolkit (for NPU), IPEX-LLM (for iGPU), PyTorch.

Deployment model: Docker containers orchestrated via RKE2 Kubernetes.

AI Models: Whisper-small (speech-to-text on NPU), LLaMA 3.2–3B (LLM on iGPU via Ollama). Testing by Celebal Technologies in February 2024.

Notices & Disclaimers

Performance varies by use, configuration and other factors.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel optimizations, for Intel compilers or other products, may not optimize to the same degree for non-Intel products.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

See our complete legal Notices and Disclaimers.

Intel is committed to respecting human rights and avoiding causing or contributing to adverse impacts on human rights. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to adverse impacts on human rights.

0925/CG/DJA/PDF Please Recycle (if printed) 367092-001US