Deloitte.

DeloitteSage

A multi-agent hybrid AI accelerator for enterprise knowledge workers

Deloitte and Intel collaborated on an advanced Multi-Agent Hybrid AI solution designed to automate routine tasks through seamless coordination of specialized AI agents. By intelligently routing on-device and cloud-based agents, this solution provides users with a unified experience, significantly enhancing efficiency and productivity.

EXECUTIVE SUMMARY

Many enterprises are reassessing their device strategy as Windows 10 sunsets, the total cost of cloud inference continues to increase, and regulators tighten data-sovereignty rules.

AI PCs powered by Intel® Core™ Ultra processors, paired with **Deloitte's multi-agent hybrid AI** architecture, provide real-time inference on-device and access to accelerated cloud resources, such as those powered by **Intel Gaudi**, per workload demands and governance requirements. The architecture cuts response latency, lowers cloud spend, and keeps sensitive data on trusted endpoints while maintaining response quality.

DeloitteSage features three agents: TechSage, MeetingBuddy and OpportunityAssist, coordinated by a LangGraph Supervisor Agent. The accelerator adapts to business scenarios, allows knowledge workers to tailor functionality, and supports rapid skill additions such as enterprise compliance checks, and research copilots.

BUSINESS CHALLENGES

Organizations eager to embed AI into everyday workflows encounter a mix of technological hurdles and rising end-user expectations that can limit adoption and reduce impact.

- **Device and application constraints**: Traditional knowledge worker tools assume constant connectivity and lack routing logic to decide where inference should run.
- **Bandwidth gaps**: Field and travelling staff lose AI assistance in low-connectivity environments.
- **Data-sovereignty pressures**: Cloud-only AI raises concerns about regulated data leaving trusted perimeters.
- **Productivity fragmentation**: Workers juggle separate apps for IT support, scheduling, and other routine tasks.
- **Seamless, familiar UX**: Users want AI features integrated into the devices and workflows they already know, without switching contexts or relearning tools.

Solution ingredients

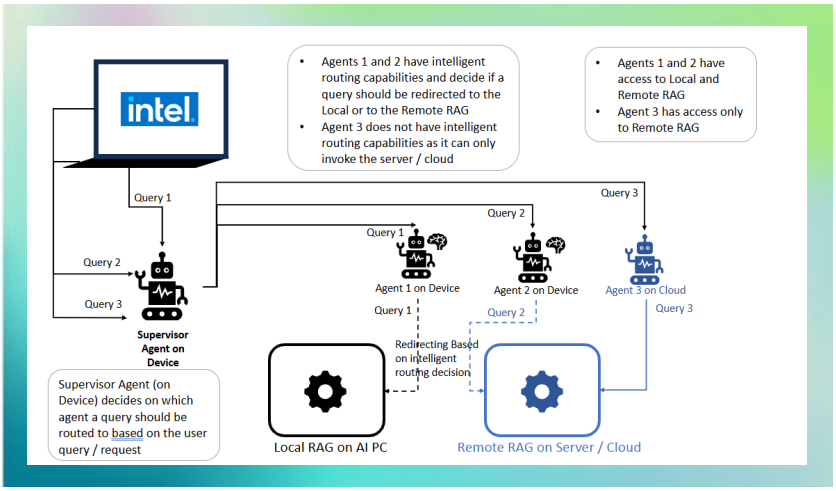
- AI PCs powered by Intel® Core™ Ultra processors
- Intel® AI for Enterprise Inference
- OpenVINO™ toolkit
- SalesForce AgentForce
- AWS EC2 Cloud Instances, pPowered by Intel M7i, etc.)
- Google Calendar API

ACCELERATOR OVERVIEW

Deloitte, in collaboration with Intel, has developed DeloitteSage, an accelerator that demonstrates the business value of a multi-agent hybrid AI approach.

Running natively on AI PCs powered by Intel® Core™ Ultra processors, DeloitteSage showcases the transformative potential for knowledge workers. Purpose-built to enable smooth routing across on-device and cloud environments, the accelerator leverages modular, intelligent agents that autonomously manage complex enterprise workflows - from IT support and meeting scheduling to CRM opportunity management, while facilitating privacy, scalability, and smooth integration with third-party ecosystems.

- **Knowledge base creation:** enables creation of knowledge based on multi-modal enterprise data for query inference
- **Multi-agent collaboration:** uses multiple specialized agents to distribute and execute steps of complex workloads
- **Centralized supervisor agent:** automatically delegates user requests to appropriate AI agents through a unified DeloitteSage interface
- **Hybrid agent deployment:** flexibility to run key hybrid agents on-device (offline) or via cloud (online), intelligently routed based on set constraints
- **Cross-agent collaboration:** supports interaction between agents, enabling end-to-end automated workflows
- **MCP tool action integration:** supports invocation of agent specific tools for performing user-approved actions
- **Privacy of sensitive data:** restricts sharing of proprietary data to the cloud without explicit user approval
- **Modular and scalable:** easy integration of additional AI agents as organizational requirements expand
- **Integration with third party agents and apps:** allows for integration with external agents and apps such as Salesforce and Google Calendar API



Each agent specializes in targeted functions, enabling comprehensive support for IT, administrative, and opportunity management related workflows.

TECHSAGE (AI IT SUPPORT)

Key functionalities

- addresses user IT queries by intelligently routing requests between local device and cloud-based services
- initiates end-to-end resolution workflows e.g., invoking Network Wizard to address Wi-Fi connectivity and related IT issues

Scope: Hybrid (on-device/on-cloud) using intelligent routine

MEETINGBUDDY (AUTOMATED SCHEDULING)

Key functionalities

- integrates directly with Google Calendar to schedule meetings based on user prompts
- summarizes meeting audio content and schedules follow-up meetings based on key discussion points

Scope: Hybrid (on-device/on-cloud) using intelligent routine

OPPORTUNITYASSIST (CRM OPPORTUNITY MANAGEMENT)

Key functionalities

- assists users in streamlining Salesforce opportunity workflows
- automates activities such as updating records and managing opportunity tags, ensuring accurate CRM management

Scope: On-cloud integration with Salesforce Agentforce

ADDRESSING CHALLENGES

To counter challenges such as device limits, patchy connectivity, data-sovereignty concerns, and rising user expectations, the accelerator keeps intelligence at the edge while scaling when more compute is required.

- **On-device acceleration:** built to run natively on AI PCs powered by Intel® Core™ Ultra processors, DeloitteSage taps both the NPU and GPU to accelerate local AI tasks and improve efficiency.
- **Adaptive hybrid routing:** hybrid agents dynamically route workloads between device and cloud based on resource checks, facilitating high-quality performance, lower costs, and data-residency compliance.
- **Offline resilience:** all on-device agents keep working in complete offline mode, safeguarding business continuity and user productivity during travel or network outages.
- **Unified work hub:** a single DeloitteSage interface can cover email, audio translation, IT fixes, and more, streamlining workflows and removing the need for multiple separate apps.

TECHNICAL ARCHITECTURE

The front-end application is built on QT C++ and has the ability to:

- start, retrieve, or delete sessions
- have a conversation with the user in text or voice formats
- provide response in text or voice formats
- regenerate the response
- start recording for a meeting and display the summary

Conversation storage and content provisioning are handled by a Python-based module that interfaces with the UI. Core functions include:

- storing all user queries and response summaries, whether generated on-device or in the cloud
- collating and supplying full conversational context to enable follow-up questions

The supervisor agent, built using LangGraph after detailed evaluation, serves as the primary backend entry point from the UI. It interprets user queries and context to determine which agent to invoke and can dynamically switch between agents within a single conversation.

Hybrid agents are architected to function both on-device and on the cloud:

- TechSage includes a retrieval-augmented generation system operating locally and on AWS, with each environment hosting its own knowledge base. It also performs device-level IT troubleshooting with user consent.
- MeetingBuddy enables on-device or cloud-based summarization via Intel's Infer-as-a-Service and can schedule standalone or follow-up meetings in Gmail based on user input and approval.

OpportunityAssist, the third agent, connects to an external platform – Salesforce Agentforce - allowing users to manage and gain insights from Salesforce data through DeloitteSage.

The decision to infer on-device or route to cloud is based on multiple common factors such as internet accessibility, AI PC resources availability (GPU, CPU, NPU, Memory) and agent-specific factors:

- complexity of user query for TechSage
- sensitivity of meeting data

Technical specs

Embedding models

- on-device: mxbai-embed-large-v1
- cloud: Amazon Titan Text Embeddings

Databases

- on-device: PGVector
- cloud: Amazon OpenSearchService

Inference models

- on-device: Meta-Llama/Llama-3.1-8B-Instruct
- cloud: Amazon Bedrock's Claude (anthropic.claude-instant-v1)

Mail service

- Gmail, integrated with MeetingBuddy Agent

Speech-to-text conversion is performed on-device using openai/whisper-tiny.en

Semantic router for dynamic agent selection: sentence-transformers/all-MiniLM-L6-v2

Text summarization

- on-device: Meta-Llama/Llama-3.1-8B-Instruct
- cloud: Intel's Infer-as-a-Service with Meta-Llama/Llama-3.1-70B-Instruct

Adaptability

This accelerator can be extended to other business use cases like:

Use case	Description
Productivity Booster	Streamlines tasks, documents, and schedules to boost efficiency
Learning & Development	Personalized learning paths with adaptive content and goal tracking
New Resource Initiation	Streamlined new-hire setup through automated workflows and verifications
Expense Report Automation	Effortless expense capture, categorization, and submission
Sales & Marketing Playbook	Automates pitches and lead insights to boost sales performance
Customer Support Modernization	Enhances support with real-time issue detection and smarter responses
Policy Compliance Analyzer	Minimizes risk through automated monitoring and proactive training
Diagnostic Support & Electronic Health Records (EHR) Summarization	Extracts key patient data and suggests triage paths to aid clinicians
Store Experience Personalization	Blends local and cloud AI for personalized, real-time customer engagement
Multi-Agent-Driven M&A Due Diligence	Automates analysis of complex data for faster, accurate deal evaluation
Manufacturing Process Optimization	Detects anomalies and performs root cause analysis to reduce downtime on the floor

LEARNINGS AND CAPABILITIES DEVELOPED

- Gained the capability to integrate agents developed on different platforms (e.g., C++ and Python) under a unified supervisor agent
- Enabled a single model instance to support multiple agents serving distinct purposes, such as Q&A and summarization
- Learned how to run models on NPUs by applying the necessary configuration changes
- Developed the ability to integrate a QT C++-based UI with multiple backend agents built in both C++ and Python
- Learned how to integrate third-party agents into the DeloitteSage ecosystem
- Gained the capability to create and invoke Model Context Protocol (MCP) action tools such as intelligent routing, Network Wizard, and Gmail calendar integration
- Developed intelligent routing capabilities to dynamically select on-device or cloud-based agents based on predefined criteria
- Built and executed a Retrieval-Augmented Generation (RAG) pipeline using a knowledge base hosted on AWS for Q&A use cases

Why multi-agent hybrid AI represents the future for businesses aiming to outpace competitors and swiftly adapt to market trends

By fusing Intel's enterprise-grade NPUs with cloud inference, DeloitteSage gives every knowledge worker a personal fleet of domain-specific agents: IT, scheduling, CRM without multiplying point solutions or exposing sensitive data to uncontrolled endpoints. Tasks that once bounced through tickets, emails, and portals are completed inside a single chat pane, cutting context-switch time and accelerating decision velocity. Because the accelerator keeps most of the routine prompts on-device and transitions smoothly to the cloud per workload demands and governance requirements, enterprises gain the latency of local execution, the depth of large models. and a predictable, policy-bound cost profile, something pure-cloud chatbots cannot match.

The accelerator runs on the modern AI-capable laptops and GPU-accelerated cloud infrastructure that most enterprises already deploy, so new capabilities arrive as a software update, not a hardware overhaul. Confidential-compute enclaves preserve data sovereignty, while a portable abstraction layer keeps the codebase ready for future processors and accelerators. Granular dashboards translate each agent's cloud utilization into clear financial impact and compliance metrics that finance and security teams can audit. As new use cases emerge, the same routing mechanism, retrieval layer, and security posture apply, giving executives a single, extensible investment that compounds in value over time.

CONCLUSION

Hybrid AI with multi-agent orchestration is more than a technical proof point. It is a strategic operating model for enterprise productivity. DeloitteSage demonstrates that when the right workload is executed on the right Intel hardware at the right moment, organizations can unlock faster answers, lower costs, and stronger governance - all from a platform that scales horizontally with new agents and vertically with advancing hardware.

In collaboration, Intel and Deloitte invite business and technology leaders to move beyond single-purpose chatbots and embrace a unified, policy-driven assistant layer that can evolve as quickly as your market, your workforce, and your imagination.

Through this collaboration, clients receive a proven launch point with complete infrastructure blueprints, enablement services and Intel's ongoing support, so they can meet business objectives and end-user demands quickly while satisfying IT mandates for security, cost control, and lifecycle management.

CONTACT US

Shakir Rizvi
AI Leader
Deloitte Consulting LLP
shrizvi@deloitte.com

Learn more about the Deloitte and Intel alliance at
https://www.deloitte.com/us/en/alliances/intel.html